Assessing Text and Web Accessibility for People with Autism Spectrum Disorder

Victoria Yaneva

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

2016

Signature: . . . . . .

Date: . . . . . .

Abstract

People with Autism Spectrum Disorder experience difficulties with reading comprehension and information processing, which affect their school performance, employability and social inclusion. The main goal of this work is to investigate new ways to evaluate and improve text and web accessibility for adults with autism.

The first stage of this research involved using eye-tracking technology and comprehension testing to collect data from a group of participants with autism and a control group of participants without autism. This series of studies resulted in the development of the ASD corpus, which is the first multimodal corpus of text and gaze data obtained from participants with and without autism.

We modelled text complexity and sentence complexity using sets of features matched to the reading difficulties people with autism experience. For document-level classification we trained a readability classifier on a generic corpus with known readability levels (*easy, medium* and *difficult*) and then used the ASD corpus to evaluate with unseen user-assessed data. For sentence-level classification, we used for the first time gaze data and comprehension testing to define a gold standard of *easy* and *difficult* sentences, which we then used as training and evaluation sets for sentence-level classification. The

results showed that both classifiers outperformed other measures of complexity and were more accurate predictors of the comprehension of people with autism.

We conducted a series of experiments evaluating *easy-to-read* documents for people with cognitive disabilities. Easy-to-read documents are written in an accessible way, following specific writing guidelines and containing both text and images. We focused mainly on the image component of these documents, a topic which has been significantly under-studied compared to the text component; we were also motivated by the fact that people with autism are very strong visual thinkers and that therefore image insertion could be a way to use their strengths in visual thinking to compensate for their difficulties in reading. We investigated the effects images in text have on attention, comprehension, memorisation and user preferences in people with autism (all of these phenomena were investigated both objectively and subjectively). The results of these experiments were synthesised in a set of guidelines for improving text accessibility for people with autism.

Finally, we evaluated the accessibility of web pages with different levels of visual complexity. We provide evidence of existing barriers to finding relevant information on web pages that people with autism face and we explore their subjective experiences with searching the web through survey questions.

Acknowledgements

Four years ago I came to the University of Wolverhampton to spend a month working on a project related to text simplification for people with autism. Little did I know that this one-month placement would be the start of a journey in academia and that today I would be writing the acknowledgments section of a PhD thesis. That placement did something even more important for me: it introduced me to the world of autism, which I am now so fascinated with.

There are so many people I have to thank: Professor Ruslan Mitkov, for encouraging me to start this PhD in the first place and for believing in me. Thank you for your faith in me, for giving me the academic freedom to follow my own ideas and for all your support;

Professor Kenneth Manktelow, for always being there for me and for showing me that statistics can actually be interesting. After each meeting with you I have felt more relaxed and confident that any problem could be solved; and

Dr. Irina Temnikova for being available in times of crises via all forms of social media in almost all hours of the day and night. You have provided me with very valuable support and advice (and occasionally with a drink, in case the first two did not work out, for which I am grateful, too!).

A huge thank you goes to Dr. Miguel Angel Rios Gaona: Miguel, I have learnt so much from you that I simply cannot imagine finishing this PhD without you by my side. Thank you for all your valuable ideas, your criticism, your code for the randomised presentation of the texts and questions, and last but far from least, for all the cooking and all the "fiestas locas", which made Wolverhampton a warmer place.

I deeply regret that, because of confidentiality, I am not able to name the dozens of participants who took part in this research. Working with you and getting to know the everyday battle some of you face coping with autism has been a transformative experience for me. It has turned this work from a mere research topic into a cause. Thank you for getting involved despite not having any incentive other than to help other people with reading difficulties. I promise I will work really, really hard to achieve this with you.

The work presented in this thesis was dependent on the involvement of the community and of many charity organisations. Special thanks in particular go to Nick Parry, Pawel Cwik and Emilia-Maria Greenwood for helping me find more and more participants during the most difficult period of my PhD. Autism West Midlands, represented by Dr. Elisabeth Hurley, Joanne Barford and Stephanie Taylor, deserve applause for their commitment, for their support of this project and for teaching me so much about autism. Others who offered me valuable help include Linda Cooper from the Goscote Greenacres Centre and the entire team of the Impact Hub in Birmingham, who helped me reach out to the public and widened the impact of this work.

I am also very grateful to all of the members of our research group. Discussions with Richard Evans and Constantin Orasan have been crucial in shaping the ideas presented in this thesis. Special thanks go to Iain Mansell for his genuine enthusiasm in helping with the outreach of this work and to Emma Franklin and George Mitkov for proofreading hundreds of pages. I met some great people in the group and, as I write these lines, I feel quite nostalgic for times past. There are too many people to mention here and I do not want to single anyone out as I have really learnt a lot from each and every one of you. You all know who you are and I am very happy that we got to share a small part of our lives together in Wolverhampton.

Finally, there are a handful of people who are not directly related to this research but who are very dear to me: My friend Sachin Sasikumar, for helping me keep a decent work-life balance and for all his songwriting; and my friends who are so far away and yet so close: Petko, Borislav, Venera, Ani, Jeni, Elena, Polina and Dobrin.

Thank you, Simeon Ivanov, for being certain that I could achieve whatever I wanted to. This belief of yours has turned into a self-fulfilling prophecy for me and, even though I am not as powerful as that, I could not have done without you and I am very glad to have known you. Thanks, too, to my sister Yana and to my parents: I will not even attempt to put into words what you mean to me. Last but extremely far from least, many special thanks go to Aaron Montgomery for... well, for putting up with me through this entire writing period (I know I am a handful) and for being the wonderful human

being that I am lucky enough to have by my side.

List of Publications

Parts of this thesis appeared in the following peer-reviewed publications:

- Yaneva, V., Mitkov, R., Orasan, C. and Temnikova, I. Can Image Insertion Aid Reading Comprehension and Memorisation in Adults with Autism? User Studies, A Prototype and Guidelines. Under Review, 2017.
- Eraslan, S., Yaneva, V., Yesilada, Y., Harper, S. and Mitkov, R. Web Users with Autism: Eye-tracking evidence of Barriers and Distractors. Under Review, 2017.
- Yaneva, V., Temnikova, I. and Mitkov, R. 2016. A Corpus of Text Data and Gaze Fixations from Autistic and Non-autistic Adults. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, Portoroz, Slovenia, 25 - 28 May
- Yaneva, V., Temnikova, I. and Mitkov, R. 2016. Evaluating the Readability of Text Simplication Output for Readers with Cognitive Disabilities. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, Portoroz, Slovenia, 25 - 28 May
- Yaneva, V., Evans, R. and Temnikova, I. 2016. Predicting Reading Difculty for Readers with Autism Spectrum Disorder. *Proceedings of Workshop on Improving Social Inclusion using NLP: Tools and*

Resources (ISI-NLP) held in conjunction with LREC 2016, Portoroz, Slovenia, 23 May

- Yaneva, V., Temnikova, I. and Mitkov, R. 2015. Accessible Texts for Autism: An Eye-Tracking Study. ASSETS 2015. The 17th International ACM SIGACCESS Conference of Computers and Accessibility, Lisbon, Portugal, 26-28 October. pp. 49-57
- Yaneva, V. and Evans, R. 2015. Six Good Predictors of Autistic Text Comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar, Bulgaria, 5-11 September 2015. pp. 697 - 706
- Yaneva, V. 2015. Easy-read Documents as a Gold Standard for Evaluation of Text Simplification Output. In *Proceedings of the Student Research Workshop at the International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar, Bulgaria, 5-11 September 2015. pp. 30-36

List of Acronyms

.

AF: Average Fixations

AI: Artificial Intelligence

ATS: Automatic Text Simplification

ATV: Average Time Viewed

AR: Average Revisits

APA: American Psychiatric Association

ARI: Army's Readability Index

ASD: Autism Spectrum Disorder

ASL: Average Sentence Length

CL: Coleman-Liau (formula)

FKGL: Flesch-Kincaid Grade Level

FRE: Flesch Reading Ease

GCSE: General Certificate of Secondary Education

GL: Grade Level

HW: Hard Words

KS: Key Stage

ID: Intellectual Disability

IQ: Intelligence Quotient

L1: First Language

L2: Second Language

LD: Language Disorders

LM: Language Model

LSA: Latent Semantic Analysis

MCQ: Multiple Choice Question

MID: Mild Intellectual Disability

ML: Machine Learning

NLP: Natural Language Processing

PCD: Pragmatic Communication Disorder

POS: Part of Speech

SLM: Statistical Language Modelling

SMOG: Simple Measure of Gobbledygook

STM: Short-term Memory

SVM: Support Vector Machine

TS: Text Simplification

WAI: Web Accessibility Initiative

WCAG: Web Content Accessibility Guidelines

WLS: Word Length in Syllables

ZPD: Zone of Proximal Development

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Having instant access to information has become crucial in our modern world, where facts, events, ideas, opinions and solutions to everyday problems are looked-up on the web with just a click of a button. Yet comprehension difficulties are a core characteristic of many people with cognitive disabilities, where merely having access to information is not sufficient as the intended message may not be fully comprehended. The term **accessibility** refers to the inclusive design of products, of services, or of environments, in such a way that everyone should be able to access them, regardless of his or her level of ability or disability. By definition accessibility must include everyone, so in this thesis we discuss accessibility of information not only as having **access to information for everyone** (by being enabled to read it) but as having **access to meaning for everyone** (by being enabled to comprehend it).

The aim of this research was to investigate new ways of evaluating and improving text and web accessibility for adults with autism. This goal was motivated by three main reasons: individuals with autism need to have access to adapted text and web content due to their reading comprehension

difficulties; the accessibility of this content needs to be robustly evaluated in order to ensure its quality meets the user requirements and finally, in order to improve text and web accessibility for this part of the population, we need to understand the barriers they encounter when reading and when using the web.

In this chapter we highlight the main challenges to improving text and web accessibility for people with autism, as well as the gaps in current research, which the experiments in this thesis address.

1.1.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised by impairment in communication and social interaction (American Psychiatric Association 2013), whose prevalence has grown rapidly from 0.5 to 14.7 per 1,000 children over the period from 1970 to 2010 (Dave & Fernandez 2015). As implied by the definition, communication difficulties are a core characteristic of those with autism. Children with autism usually acquire language later in their lives compared to peers without autism (Frith & Snowling 1983). This delay in language development results in language comprehension and reading comprehension difficulties such as:

- resolving ambiguity in meaning (Happé & Frith 2006, Happe 1997, Frith & Snowling 1983, O'Connor & Klein 2004, Martos et al. 2013);
- syntax processing of long sentences (Whyte et al. 2014);
- identifying pronoun referents (O'Connor & Klein 2004);

- having difficulties in figurative language comprehension (MacKay & Shaw 2004);
- making pragmatic inferences (Norbury 2014).

In addition to atypical text processing among some individuals with autism, there are also differences in attention span between autistic and non-autistic people (Lovaas & Schreibman 1971), which have been shown to affect reading behaviour (Happé & Frith 2006).

Social media and the web are particularly important to people with autism because they allow them to connect to other people without being impeded by the complexities of social situations, which they find particularly challenging (Bosseler & Massaro 2003). Evidence for the need of people with autism to have access to autism-friendly social media is provided by the development of platforms such as the UK-wide Autism Connect[1]. Autism Connect allows users with autism to connect to each other in a moderated accessible environment, where the rules and the means of navigation are explained using easy-to-read language at every step of the process:

> "*Account - an account is a record of your details. Every user has an account that they have to log in to. The account remembers the things you do and the things that other people have said and done in reply to you* [2] "

The characteristics of autism and the way they affect reading comprehension and web searching are discussed in detail in Chapter 2.

[1] Autism Connect. https://www.autism-connect.org.uk/

[2] Autism Connect - Site Use. https://www.autism-connect.org.uk/index.php/sitesiteuse

1.1.2 Text Accessibility

As we shall see in Chapter 2, people with autism experience pragmatic difficulties rather than difficulties with word decoding. For this reason, we focus on improving ease of comprehension through investigating text complexity and strategies to aid comprehension, rather than improving text legibility (e.g. through investigating font sizes and types).

One way to enhance the reading comprehension of individuals with cognitive disabilities is to provide them with accessible texts, that do not contain linguistic constructions that may be challenging for the target population (e.g. long sentences, passive voice, figurative language). There are a number of initiatives that promote accessible writing such as the *Plain English campaign*[3] and the *Easy To Read campaign* (Tronbacke 1997). The aim of these initiatives is to produce 'easy-to-read' documents: accessible documents produced by humans, following a set of writing guidelines such as the European "Make It Simple" guidelines (Freyhoff et al. 1998) or 'Guidelines for Easy-to-read Materials' (Nomura et al. 2010). Governmental and healthcare organisations in the UK and the USA are required to produce accessible versions of important documents by the UK Equality Act 2010[4] and the Americans with Disabilities Act[5], respectively. Having easy-to-read versions of important documents is also a practice in many charity organisations such

[3]Plain English Campaign. Available at: http://www.plainenglish.co.uk/

[4]Equality Act 2010. UK. Avaliable at: http://www.legislation.gov.uk/ukpga/2010/15/section/6.

[5]Americans with Disabilties ACT Available at: http://www.ada.gov/cguide.htmanchor62335

as Britain's National Autistic Society[6].

Easy-to-read documents have two main components:

Text: The text used in easy-to-read documents does not contain long sentences, complicated words or linguistic constructions which may be challenging for the specific target group and is produced following specific guidelines (e.g. the ones proposed by Freyhoff et al. (1998)).

Images: Images in easy-to-read documents complement and reinforce the text. Currently, there are only very limited guidelines for the choice of images in easy-to-read documents. Even though people with autism often have difficulties inferring meaning from symbols and drawings as opposed to photographs (Sampath 2010), currently both types of images are widely used in easy-to-read documents (Chapter 6). Furthermore, no information exists on autistic adults' preferences regarding images in texts.

Figure 1.1 shows an extract from an easy-to-read document: an accessible version of the UK Conservative Party's manifesto for the 2015 general election.

Even though demand for easy-to-read documents is high, writing and evaluating them is time-consuming and expensive. In an attempt to solve this problem, campaigns to educate and train professionals to produce easy-to-read information (particularly for healthcare documents) (Plimpton & Root 1994, Root & Stableford 1999) are underway. These seek to address "the mismatch between the low literacy skills of the target population and the

[6]National Autistic Society website: http://www.autism.org.uk/

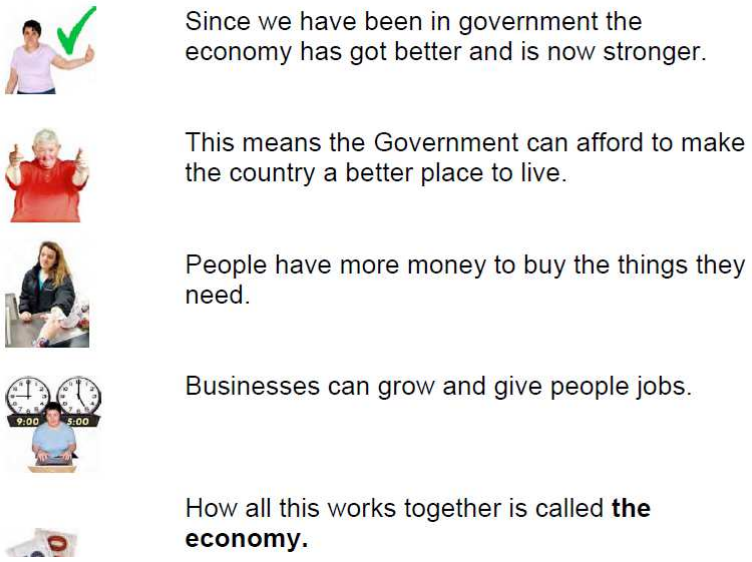

Figure 1.1: Conservative Party Manifesto Easy-to-Read Version (extract) (2015)

high reading level of most health and managed care materials" (Root & Stableford 1999).

Thus, not only the quality of published easy-to-read materials but also the efficiency of the development of such content has become a key area for improvement in the campaign for accessible information.

Automatic Text Simplification (ATS) is an application in the field of Natural Language Processing (NLP), the main task of which is to convert texts into a more understandable form for readers with lower than average reading skills, without changing the original meaning of the text. Unfortunately, ATS technology is not yet mature enough to adapt a document fully automatically to meet the needs of readers with autism, which is why these

documents are still developed mainly by humans; however, NLP has the potential to aid certain stages of the production of accessible texts. One such stage is the evaluation of these documents, which could be aided through automatic readability assessment.

1.1.3 Web Accessibility

The Web Accessibility Initiative was launched in 1997 by the World Wide Web Consortium (W3C)[7]; it aims to "lead the web to its full potential to be accessible, enabling people with disabilities to participate equally on the web"[8]. Since 1997, there have been remarkable advances in enabling people with motor, visual or hearing impairments to use the web, such as the development of speech-to-text and text-to-speech converters or the development of accessible .html templates for screen readers. As inspiring as these advances are, accessibility development for individuals with cognitive disabilities seems to be lagging behind (Friedman & Bryen 2007) and cognitive issues have been assigned lower priorities in the web accessibility guidelines (Britto & Pizzolato 2016). One possible reason for this lag is the fact that cognitive disabilities such as autism, intellectual disability, dyslexia and hyperactivity are "hidden" disabilities, meaning that there appear to be no web-accessibility barriers for those who have such disabilities, while in real-

[7]World Wide Web Consortium (W3C) International Web Accessibility Initiative [online] Available at: https://www.w3.org/Press/WAI-Launch.html

[8]W3C Accessibility. Available at: https://www.w3.org/standards/webdesign/accessibility

ity this may not always be the case. Cognitive disabilities have an impact on the way people use the web, owing to "limited reading comprehension, complexity, slower learning, limited fine motor control (...) and lowered information overload threshold" (Friedman & Bryen 2007).

Among the various web accessibility sets of guidelines in existence, the most widely used are WCAG 2.0 created by W3C/WAI (Caldwell et al. 2008), which aim to meet the needs of all disabled user groups. Unfortunately, issues related to cognitive disabilities are the ones least discussed in WCAG (Harper & Yesilada 2008). There are a number of web accessibility sets of guidelines developed specifically for people with autism; however, as shown in a comprehensive review by Britto & Pizzolato (2016), most of them have not been empirically tested and validated through scientific studies.

1.1.4 Problems with Evaluating Text and Web Accessibility

In cases where the easy-to-read content is targeted at people with cognitive disabilities, accessible-content manuals mandate that the output text be evaluated by a focus groups of target users. In spite of the fact that these guidelines recommend the use of readability formulae for assessing the complexity of the final document, the European guidelines for writing easy-to-read documents state the following:

> "*To ensure that your document really meets the needs of your target group and is suitable for their reading abilities, it is essential*

> *that people with learning disability or groups of self-advocates read it before it is printed. This is the only way to ensure your publication really meets the needs and abilities of your target group, thus increasing the number of potential readers.(Freyhoff et al. 1998)* "

However, recruiting a focus group of people with cognitive disabilities is not straightforward. The main issues lie in the difficulty of recruiting enough participants with the required profile, the fact that this type of evaluation is particularly time-consuming, and the substantial funding required. Furthermore, not all people with cognitive disabilities have the same reading difficulties and even if access to such a group is secured, it is not feasible for the participants to evaluate lengthy documents (Brega et al. 2015). All of these barriers prevent the robust evaluation of the easy-to-read materials currently being produced and most developers of easy-to-read materials do not report any evaluation for the needs of people with cognitive disabilities.

Another aspect of easy-to-read documents is the images used; however, this aspect has almost not been investigated. It is currently not known how exactly images affect comprehension and memorisation in people with different types of cognitive disabilities; various types of images are being used (both suitable and unsuitable) as there are no comprehensive guidelines regarding the image type; positioning of the image also varies widely, with most of the images being placed either on the left or the right hand-side of the paragraph; last but not least, little is known about the user preferences of the different groups of readers. As a result of this lack of clarity regarding the use

of images, there are no robust procedures for the evaluation of this component of the documents. This problem is discussed in more detail in Chapter 6, where we focus on the image component of the easy-to-read documents.

In addition to having to produce accessible text content, making the web accessible for people with cognitive disabilities entails investigation of a number of design and interaction issues. There are almost no studies investigating the way people with autism interact with the web and there is a lack of understanding about the way they process web pages, the way visual content affects their attention and, last but not least, what their user preferences are.

To summarise, there are currently three main barriers to making text documents and the web pages accessible to individuals with autism:

1. The development and evaluation of accessible texts is time-consuming and costly.
2. There is lack of understanding of how the image component in easy-to-read documents affects text comprehension and memorisation among people with autism. Furthermore, there is no evidence to suggest which types of images (e.g. photographs, drawings, symbols) are most suitable to be used.
3. There is a lack of understanding about the way people with autism interact with the web and about how to improve the web accessibility guidelines for this part of the population.

1.2 Goals

The goal of this thesis is **to investigate new ways to evaluate and improve text and web accessibility for adults with autism**.

This primary goal encompasses several secondary goals, which correspond to the research questions (RQs) outlined in the next section:

1. **To evaluate automatically the accessibility of text content for readers with autism (RQ1, RQ2, RQ3)**

   The accessibility of text content is known as text readability. Readability has been defined as the ease of comprehension deriving from the style of writing (Harris & Hodges 1995) (Chapter 2). We will evaluate content readability at **document level** (Chapter 4) and **sentence level** (Chapter 5), through the training and evaluation of classifiers based on machine learning (ML) algorithms.

2. **To investigate the effects of the presence of images and the types of images on text comprehension and memorisation among readers with autism (RQ4)**

   A number of accounts show that people with autism have a strong preference for processing visual over verbal information (Kana et al. 2006, Grandin 2009, Quill 1997, Dettmer et al. 2000). We will aim to find out whether their ability to process visual information could be used to compensate for their reading comprehension difficulties and to improve guidelines for writing easy-to-read material by investigating

the role of images in these documents.

3. **To investigate the accessibility of web pages for those with autism (RQ5)**

   Since the Internet is nowadays one of the main sources of information, we will aim to ascertain whether those with autism encounter barriers when searching for information on the web and, if so, what these barriers are.

1.3 Research Questions

We will achieve our aforementioned goals by answering the following research questions (RQ).

RQ1: How can we obtain a collection of texts with known levels of difficulty for readers with autism?

Motivation: Currently, there are no existing corpora containing texts with known levels of difficulty for people with autism; therefore, the evaluation of a readability classifier would not be feasible without the compilation of such a corpus.

Method: We conducted a series of reading comprehension experiments involving participants with and without autism. We used text comprehension questions and eye-tracking data to determine the level of difficulty of text

passages (*easy, medium* or *difficult*) for readers with autism.

Result: The result of the reading comprehension experiments is the compilation of the ASD corpus, which we used for the evaluation of our readability classifiers. A detailed description of the compilation of the corpus is provided in Chapter 3.

RQ2: Is it possible to develop an automatic *document*-level readability classifier for people with autism, that generalises over unseen user-evaluated data better than existing readability metrics?

Motivation: As explained above, there is a need to develop a robust method to evaluate text readability for people with autism; one which does not depend on access to a focus group of participants, which allows the evaluation of lengthy documents and which is more accurate than existing readability metrics.

Method: We modelled text complexity using a set of linguistic features related to the reading difficulties of people with ASD and used supervised machine learning algorithms to train a document-level readability classifier. We evaluated its internal validity by means of cross-validation and then evaluated its capacity to generalise over the unseen, user-evaluated ASD corpus.

Result: The result of these experiments is the development of a document-level readability classifier, which outperformed a common baseline. A detailed description of the classifier is provided in Chapter 4.

RQ3: Is it possible to develop an automatic *sentence*-level readability classifier for people with autism, that performs better than existing readability metrics?

Motivation: Evaluation of sentence-level readability for people with autism can indicate particular sentences in the text which pose problems for the reader. This assessment is particularly useful as an ad-hoc step for automatic text simplification systems because it allows them to identify and simplify only the complex sentences in a text, while leaving the rest of the text intact.

Method: We used the number of eye gaze fixations for each sentence of the ASD corpus as a measure of sentence complexity (a higher number of gaze fixations is indicative of higher complexity (Chapter 2)). We evaluated the complexity of an additional set of 100 sentences by using comprehension testing involving participants with autism. The combined dataset was used for the training and evaluation of the sentence-level readability classifier.

Result: The result of these experiments is the development of a sentence-level readability classifier, which outperformed a common baseline. A detailed description of the classifier is provided in Chapter 5.

RQ4: Do images inserted into texts have an effect on participants' attention, comprehension and memorisation of a text, measured both objectively and subjectively?

Motivation: The motivation for investigating the role of images is two-fold. First, images are an inherent part of easy-to-read documents but they are only very rarely mentioned in the official writing guidelines. Second, people with autism are very strong visual thinkers (Chapters 2 and 6), hence relying on their strengths in using visual information has the potential to compensate for the reading comprehension difficulties they have. There is currently no clarity regarding the types of images that should be used, how they need to be positioned within the text or how they affect comprehension and memorisation among readers with autism.

Method: We conducted a number of comprehension and eye tracking experiments to investigate the effects of images on attention, comprehension and memorisation among people with autism (measured objectively and subjectively); user preferences regarding the inclusion of visual cues within the text; and user preferences regarding the positioning of images.

Result: The results of these experiments are synthesised in detailed accessibility guidelines regarding images and easy-to-read documents for users with autism (Chapter 6).

RQ5: Do web users with autism encounter barriers to finding information on web pages?

Motivation: The web has become one of the major sources of information in the modern world. Although people with cognitive disabilities are known to experience difficulties using the web, there is currently no evidence to confirm whether people with autism experience any barriers when looking for information within web pages and, if so, what these barriers are and how they could be removed.

Method: We conducted an experiment in which we compared the success rates of people with autism in finding specific information on web pages to the success rates of participants without autism. We collected eye-tracking data in the process of their searching in order to determine whether differences exist between the cognitive effort of the two groups and asked survey questions in order to explore what other difficulties with web searching they might have. We also investigated the effect the visual complexity of the web pages has on the success of finding the required information.

Result: The results indicated that the participants with autism found it significantly more difficult to search for information within web pages (Chapter 7).

1.4 Original Contributions

The investigation of the research questions outlined above lead to the following original contributions of this thesis. It is important to note that, for the purpose of clarity, the original contributions have been listed based on the order in which they will appear in the thesis and not based on their strength and originality.

Contribution 1. The development of a corpus of texts with known levels of reading difficulty for readers with autism (the ASD corpus) (RQ1)

The ASD corpus is the first text collection whose difficulty has been evaluated by readers with autism. Furthermore, the ASD corpus is the first text collection which contains gaze data from people with autism. In the context of our research, the ASD corpus was developed to serve as a test set of unseen data for our readability classifier. However, the annotation of the corpus and the gaze data it contains for each word from both sets of participants (from those with and those without autism), make it a suitable resource for investigating reading differences between these groups. To the best of our knowledge, this is the first time gaze data have been used to study reading among people with autism; thus the ASD corpus is a valuable resource not only for the NLP community but also for anyone interested in psycholinguistic investigation into reading among people with autism.

Contribution 2. The development and evaluation of a *document*-level readability classifier for readers with autism. (RQ2)

The document-level readability classifier described in Chapter 4 is the first automatic readability classifier for people with autism. Furthermore, it is the first readability metric for this group to be evaluated on data obtained from participants with autism. In Chapter 4 we will demonstrate that the classifier outperforms the well-known Flesch-Kincaid readability formula (Kincaid et al. 1975) by achieving 77% accuracy in distinguishing between *easy, medium* and *difficult* texts when evaluated on unseen data.

Contribution 3. The development and evaluation of a *sentence*-level readability classifier for readers with autism. (RQ3)

The sentence-level readability classifier developed in this thesis is the first ever developed for people with autism. It is important to note that all previous studies on sentence-level readability assessment use manual simplification or rankings by experts as a gold standard of text accessibility (Inui et al. 2001, Vajjala & Meurers 2014, Pilán et al. 2014). In this thesis we will present a different approach, in which our gold standard is based on the eye tracking data obtained from our participants with autism, as well as their comprehension of sentences in a vocabulary test (Chapter 5). Furthermore, all sentences in our training and test data are naturally occurring sentences as opposed to simplified versions of other sentences, where sentence length and lexical complexity have been manipulated and thus may have been skewed by bias.

Our sentence-level classifier outperformed the baseline measure of sentence length by achieving 81% accuracy for 10-fold cross validation.

Contribution 4. Improved text and web accessibility guidelines for people with autism. (RQ4)

The results of eye-tracking and reading comprehension experiments were synthesised in the form of guidelines. These guidelines are so far the most detailed ones with regard to the use of images in text, outlining the optimal image type and image positioning. The guidelines also contain recommendations for improving comprehension, memorisation and for coping with slower reading speed among readers with autism.

Contribution 5. Evidence for barriers encountered by individuals with autism when they search for information within web pages. (RQ5)

These findings are the first step towards filling in an existing gap in accessibility research, namely the gap between the needs of people with autism and the way web content is organised and presented. We will show that web users with autism experience much greater difficulty finding information on web pages compared to the control group. We also explore the effects of the visual complexity of the web pages on finding relevant information and compare the cognitive load the tasks imposed on the two groups of participants. Finally, we explored the participants' experiences with web searching by using survey questions.

1.5 Structure of the Thesis

The rest of this thesis is organised as follows.

- **Chapter 2** presents related work in the fields of readability research, autism and the use of eye tracking techniques for the investigation of text complexity.

  We then discuss ways to automatically evaluate the difficulty of text content for readers with autism:

- **Chapter 3** describes the development of the ASD corpus, which is used for the evaluation of the readability classifiers. This chapter refers to RQ1 and Original Contribution (OC) 1.
- **Chapter 4** describes the training and evaluation of a document-level readability classifier for people with autism. This chapter refers to RQ2 and OC2.
- **Chapter 5** describes the training and evaluation of a sentence-level readability classifier for people with autism, as well as the way gaze data and comprehension testing at sentence level were used for the development of a gold standard of *easy* and *difficult* sentences. This chapter refers to RQ3 and OC3.

  Then, we discuss the role of images in easy-to-read documents:

- **Chapter 6** describes reading-comprehension and eye-tracking experiments examining the use of images in text, as well as detailed guidelines for their use. This chapter refers to RQ4 and OC4.

Finally, we discuss the accessibility of web pages:

- **Chapter 7** describes web-search and eye-tracking experiments examining how people with autism locate relevant information on web pages. This chapter refers to RQ5 and OC5.

  The main findings on text readability, images in text and web-page accessibility are discussed in the last chapter:

- **Chapter 8** revisits the main research questions and original contributions of this thesis, comments on their strengths and limitations and proposes avenues for future research.

Chapter 2

Background

2.1 Chapter Overview

This chapter presents background studies on the topics of 1) reading among those with autism, 2) readability research and 3) the state-of-the-art in eye-tracking for investigation of text complexity.

2.2 Autism Spectrum Disorder

This section presents the main characteristics of autism as well as the reading difficulties it entails.

2.2.1 Main Characteristics

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised by impairment in communication and social interaction (American Psychiatric Association 2013). More concretely, the formal diagnostic criteria proposed by the American Psychiatric Association (2013) are as follows:

1. Persistent deficits in social communication and social interaction across multiple contexts.

2. Restricted, repetitive patterns of behaviour, interests, or activities.

3. Symptoms must be present in early childhood (but may not become fully manifest until social demands exceed limited capacities, or may be masked by learned strategies later in life).

4. Symptoms cause clinically significant impairment in social, occupational, or other important areas of current functioning.

5. Not better explained by intellectual disability or global developmental delay.

One of the first signs of the qualitative impairment in verbal and non-verbal communication is language delay. More often than not, children with ASD start speaking around the age of five, with some of them remaining non-verbal throughout their whole lives. Others, mainly with Asperger's syndrome, do not experience language delay (Frith 2003). While some people on the autism spectrum (usually the ones with Asperger's syndrome) may not experience reading-comprehension difficulties, others may find a range of linguistic phenomena challenging to use and comprehend.

Some verbal children on the autism spectrum may have difficulties using pronouns properly (until about 6 years of age) and sometimes they tend to use phrases not typical for their age (in some cases, even more advanced phrases than expected), idiosyncratic speech and neologisms (Oliver 1998). Echolalia (compulsive repeating of a phrase) and fixation on a certain topic are typical, as well as deficits in understanding intonation or sign language. Pragmatic components of communication, such as understanding ambiguity,

figurative language, humour, sarcasm, etc. may also be impaired.

In terms of working memory, which is important for reading comprehension, some people with ASD are found to have deficits, while others perform within the normal span. Bennetto et al. (1996) found that individuals with high-functioning autism performed worse than control subjects on a task on working memory and temporal-order memory, but not on short- and long-term recognition, cued recall, or new learning ability. Russell et al. (1996) confirm that the working memory of children with autism is limited compared to the working memory of the control subjects and functions in a similar way to that of children with moderate learning abilities. Williams et al. (2005) report that there was no deficit in verbal working memory, but they found a slight impairment in spatial working memory in individuals with autism. Overall, although not totally impaired, working memory in autism functions differently from working memory in non-autistic subjects, which may have implications for the processing of long sentences, resolving anaphora and following the discourse of a text.

Among all subtypes of ASD, the ones relevant to our work on readability assessment are the ones where reading abilities have been developed and intellectual abilities are intact (there is no developmental delay). Subtypes which would meet this criteria would be high-functioning autism (IQ level above 70), Asperger's syndrome, Pragmatic Communication Disorder, and, to some extent Pervasive Developmental Disorder Not Otherwise Specified. Individual cases from other subtypes may also occasionally share these char-

acteristics.

The next section describes the particular reading difficulties which some people with autism experience. At the end of each subsection we propose possible linguistic features which could be used to account for the specific reading difficulty discussed. Later, in Chapters 4 and 5 we present these features in detail in the context of the document-level and sentence-level readability classifiers for people with autism.

2.2.2 Reading Comprehension in People with Autism

While there is no unanimous explanation for the pragmatic deficits in reading comprehension of readers with autism, possible reasons could be their preoccupation with text and reading, the tendency to be interested in local features rather than global coherence, or a particular cognitive pattern (Nation et al. 2006).

Before discussing the character of the various linguistic phenomena which cause reading difficulty for ASD readers, we present the results of an empirical study by Martos et al. (2013), involving 120 participants from UK, Spain and Bulgaria. The participants were both children (12-18) and adults (18+), who were presented with reading tests containing various tasks evaluated through multiple choice questions. The reading tests were specifically designed to explore which linguistic phenomena are difficult for readers with autism. The authors ranked the results in categories of high and medium priority. The results are shown in Table 2.1.

Table 2.1: Text properties with high level of difficulty for readers with autism (Martos et al. 2013)

Language	**Priority**
Metaphors	High
Literal vs. Inferred meaning	High
Polysemy	High
Less common words	High
Phraseological units	High
Long words	High
Internet slang	High
Verbs expressing mental states	High
Adjectives expressing emotions	High
Subordinate adjective clauses	High
Consecutive and concessive clauses	High
Infinitive and gerund clauses (except when)	High
Adversative conjunction (except when)	High
Complicated subordinate conjunctions	High
Adversative conjunction	High
More than one clause per sentence	High
Negative and double negative sentences	High
Anaphors	High
Inferences	High
Long paragraphs	Medium
Less common ortographic signs (&, %, /)	Medium
Acronyms and abbreviations	Medium
Improper grammar	Medium
Adjectives of nationality	Medium
Words cut at the end of line.	Medium
Adverbs ending with the suffix “ly”	Medium
Infrequent conjunctions and prepositions	Medium
Passive voice	Medium
Paragraphs cut at the end of the page	Medium
Temporal adjectives	Low

Although the authors do not analyse the reasons for these results, the classification they propose is a valuable resource for the task of readability assessment for people with ASD. The above findings accord with other studies focusing in more detail on one or several of these phenomena. The atypical cognitive processes which turn these linguistic phenomena into obstacles to reading comprehension among people with autism are discussed below.

2.2.2.1 Word decoding and reading accuracy

In order to comprehend the intended meaning of a text, the reader has to be able to decipher the words using the phonological loop of the working memory and afterwards to apply various language-processing mechanisms to access the overall pragmatic intention of the text (Perfetti et al. 2005). These two stages correspond to two processes performed during reading: decoding and comprehension. Since people with ASD have certain difficulties processing written text, it is worth exploring whether these difficulties occur at the level of decoding or at the level of comprehension.

Nation et al. (2006) point out the large number of case studies which report exceptional levels of reading skill in people with autism. Such advanced ability to recognise written words often appears in the context of developmental disorders and is called hyperlexia (Frith & Snowling 1983). Hyperlexia is characterised, especially in autism, by a preoccupation with word decoding as an activity, rather than as a means to access information. Indeed, hyperlexia often entails great difficulty in capturing the gist of a text. Common in

hyperlexia are the early onset of word recognition and the general mismatch between reading accuracy and cognitive and social deficits (Frith & Snowling 1983).

Frith & Snowling (1983) explore whether people with autism can make use of both lexical strategy for reading familiar words (look-and-say) and phonological strategy for reading unfamiliar ones, based on grapheme-to-phoneme conversion. The results suggest that phonological processing in readers with autism is intact and thus decoding is not the source of the problem. Other observations made by the authors include that it may be difficult for people with autism to integrate the semantic meaning of a word into their world knowledge, gained through experience (Frith & Snowling 1983). This would in turn cause impairment in their ability to disambiguate words. In most cases the subjects knew the meaning of a single word but could not make use of context to understand its role in the text, which evoked the conclusion that "their particular problem in comprehension lies not within the 'inner lexicon', but within the 'inner encyclopaedia'" (Frith & Snowling 1983).

Nation et al. (2006) argue that Frith and Snowling's results are due to the small span of their sample: eight autistic children with IQ range 54-103 (mild intellectual disability to high-functioning). Exploring the same problem with a larger and more diverse sample of 41 children with ASD with various IQ scores, Nation et al. (2006) come to the following conclusion: "Generally, our data demonstrate rather low levels of non-word reading ability, and for

some individuals, non-word reading skills were considerably below the level expected given their level of word reading ability" (Nation et al. 2006). They also recommend that "it is important for both research and practice that the heterogeneous pattern of reading skills in children with autism is recognised" (Nation et al. 2006).

The conclusion that decoding ability in readers with autism is intact in many cases of hyperlexia but varies in non-hyperlexic autistic readers is valuable for the task of readability assessment. First, it rules out reading time as a possible measure of text complexity for this population. Second, it suggests that autistic readers with hyperlexia, who do not have dyslexia as a co-morbid disorder, would not benefit much from readability metrics designed for dyslexic people and vice-versa. On the other hand, autistic readers who are not hyperlexic have some reading difficulties in common with people with dyslexia.

2.2.2.2 Resolving anaphora

Anaphora is defined as a linguistic phenomenon of pointing back to a previously mentioned item in the text, where the "pointing back" is called *anaphor* and the entity to which it refers is its *antecedent* (Mitkov 2002). Anaphoric devices are pragmatic signals that inform listeners or readers where to search for a referent (O'Connor & Klein 2004). The most common anaphoric device is the pronoun, which points back to a recently discussed referent, the knowledge of which is still stored in the working memory. O'Connor & Klein (2004)

report that less-skilled readers with ASD make errors in identifying pronoun referents, and that these errors become more common with the complexity of the sentences.

These errors may be due to the potential working memory deficits of readers with autism as they do not store information for the antecedent or cannot access this information. Another possible explanation for their poor anaphora-resolutions skills is the Theory of Mind (Hadwin et al. 1997). Theory of Mind explains their difficulties in understanding pronouns with the overall difficulty in understanding subjectivity and the different roles of people, which pronouns signify (e.g. when one person speaks, he or she refers to him/herself with the pronoun "I" but other people address the same person with "you"). Because of the same impaired ability to understand that pronouns depend on the person speaking, some children with autism may refer to themselves using their proper name instead of "I" or simply speak of themselves the way other people address them (using "you"), until about six years of age (Oliver 1998).

Readability features which could be used to measure the difficulties anaphora resolution poses to readers with autism are *noun, argument, stem* or *anaphor overlap, pronoun incidence* and *lexical chains* among others. These features are defined and described in detail in Chapters 4 and 5, as well as in Section 2.3.4.

2.2.2.3 Resolving ambiguity in meaning

Frith & Snowling (1983) report that readers with autism have a good understanding of syntactic context, meaning that in a cloze test task they pick syntactically well-matched words, but they have an impaired understanding of semantic context, as these words are syntactically appropriate but semantically inappropriate. Due to the deficit in understanding semantic context, some readers with autism with low verbal ability perform poorly on disambiguating homophones (words which are pronounced the same way (and in many cases spelled the same way) but have different meanings) (Frith & Snowling 1983, Happe 1997, O'Connor & Klein 2004).

Prior knowledge also plays a significant role in the process of disambiguation. However, readers with autism are known to make limited use of prior knowledge even on topics they are familiar with, which directly affects word disambiguation and integration of text above the level of the clause (O'Connor & Klein 2004). Strategies, facilitating reading comprehension, such as activation of prior knowledge through pre-reading questions, are shown not to enhance comprehension significantly, and in some cases even to have negative effects, as activation of irrelevant or inaccurate prior knowledge causes confusion (O'Connor & Klein 2004).

Examples of readability features which could be used to measure ambiguity in meaning are *Number of polysemous words* and *Polysemous type ratio*, described in detail in Chapter 4.

2.2.2.4 Making pragmatic inferences

Day & Park (2005) describe the process of making pragmatic inferences while reading as the ability to combine two pieces of implicit information in order to arrive at a third piece of information, which is also implicit. Dennis et al. (2001) investigated the ability of readers with high-functioning autism to make pragmatic inferences from implicit information in the text. More concretely, they investigate:

1. Pragmatic inferences about given or presupposed knowledge in mental-state words
2. Pragmatic inferences about new or implied knowledge in mental-state words
3. Bridging inferences essential for coherence
4. Figurative language
5. Speaker's intention

The subjects managed to explain the meanings of words and to understand when a word is polysemous, as well as when to use given or presupposed knowledge to infer the mental states of characters. However, what turned out to be challenging for them was interpreting the intention of the author, examples of metaphor and comprehending what mental-state words imply in context (Dennis et al. 2001).

Readability features which could potentially be used to measure the number of inferences required from the reader to comprehend the text are features of cohesion. These include *Number of temporal conjunctions* or *Causal conjunctions*, Incidence scores of *Pronouns* and *Definite descriptions*, *Number of illative conjunctions*, Comparative conjunctions, *Adversative conjunctions*, etc. These features are defined and described in detail in Chapter 4.

2.2.2.5 Comprehending comparisons and similes

Comparisons are phrases that express the similarity of two entities. They rely on specific patterns that make them recognisable: "be like", "be as", "as as". Similes are a subset of comparisons. A simile is a figure of speech that builds on a comparison in order to convey certain attributes of an entity in a striking manner (e.g. "Love is like a flame").

People with ASD show almost no impairment in comprehending those similes which have a literal meaning (Happé 1995). This relative ease in processing is probably due to the fact that similes contain explicit markers (e.g. "like" and "as"), which evoke comparison between two things in a particular way.

With regard to understanding figurative similes, Hobson (2012) describes in the case of fifteen-year-old L.: "He could neither grasp nor formulate similarities, differences or absurdities, nor could he understand metaphor" (Hobson 2012).

Theoretically, one of the most obvious markers of similes, the word "like",

could be a source of a lot of misinterpretations. For example, like could be a verb, a noun, or a preposition, depending on the context. Given that people with autism have problems understanding context, it is interesting to consider how an autistic reader would perceive the role of "like" in a more elaborate and ambiguous comparison.

Work towards simile recognition for the purposes of readability assessment and text simplification for people with autism is described in Niculae & Yaneva (2013).

2.2.2.6 Comprehending figurative language

One of the main characteristics of language use among people with autism is literalness. Not only it is atypical for autistic people to use figurative phrases, but some of them also have severe difficulties comprehending them. Happé (1995) describes:

> *"A request to "Stick your coat down over there" is met by a serious request for glue. Ask if she will "give you a hand", and she will answer that she needs to keep both hands and cannot cut one off to give to you. Tell him that his sister is "crying her eyes out" and he will look anxiously on the floor for her eye-balls..."*

Oliver (1998) describes a general inability to tell when something is literal, when it is figurative or when it is not true at all. People with autism may adopt the idea that what they read or hear sometimes means a different thing from what they think it does, but this uncertainty is not enough to help them decipher the intended meaning.

Some studies indicate that people with ASD can learn certain idiomatic expressions and use them appropriately in social situations (Whyte et al. 2013). With the exception of some idiomatic phrases and conventional metaphors, figurative language and the speaker's intention remain, largely, areas of the utmost difficulty for autistic readers. MacKay & Shaw (2004) conduct a comparative study of the comprehension of different tropes and the pragmatic intention behind them by people with and without autism. The investigated figures of speech are:

- Hyperbole (exaggeration)
- Indirect request (e.g. "These potatoes look delicious"; important for testing comprehension of speaker's intention)
- Irony (a rhetorical device expressing a contrast between what is being said and what the truth, or the intention or the attitude of the speaker is)
- Metonymy (using an associated meaning instead an object or concept's actual name)
- Rhetorical questions (questions implying an answer)
- Litotes (underestimation)

The results suggest that people with autism struggle more to identify the pragmatic intention of a phrase and less to comprehend its meaning. Phrases,

referring to relative time or quantity, such as "it did not take me long" and "just a few things" turned out to be hardest to comprehend. The subjects tried to guess the exact time or quantity, did not understand the concept of relativity, nor made use of context to define the concrete meaning. Just a few subjects managed to recognise the "untruthfulness" of hyperbole and litotes but hyperbole was considered more of a means to impress others or to "show off".

Indirect requests and rhetorical questions were successfully understood both in terms of meaning and intention, unlike other figures of speech. Among all tropes, metonymy and irony made comprehension the most difficult. The results of the experiment suggest that the relative difficulty of the investigated tropes ranks as follows:

1. Metonymy and irony
2. Hyperbole and litotes
3. Indirect request and rhetorical questions

The study by MacKay & Shaw (2004) does not discuss the relative difficulty of metaphor, which is considered to be the most elaborate figure of speech from the cognitive point of view. Lakoff & Johnson (2003) point out that while the process of comprehending metonymy involves mapping concepts from the same semantic domain (there is a strong association between the two concepts), metaphor requires more elaborate cross-domain mapping.

Rundblad & Annaz (2010) compare ASD readers' comprehension of metaphor and metonymy and come to two major conclusions:

- Readers with autism show much lower comprehension of metaphor than of metonymy
- Metaphor comprehension does not improve with age

Rundblad & Annaz (2010) establish metaphor as a phenomenon of the highest level of difficulty for readers with autism, the understanding of which does not improve or deteriorate with age.

While metaphor is a figure of speech that is highly challenging for most people on the autism spectrum, its computational recognition is not a trivial task (Shutova 2010, Mohler et al. 2014, Hovy et al. 2013). As a result of this challenge, features relating to the automatic detection of figurative language have not been explored in the development of the readability metrics presented in this thesis.

2.2.2.7 Summary of the reading difficulties of people with autism

The main impairment related to reading comprehension in autism is pragmatic impairment. Reading difficulties experienced by people on the autism spectrum are mainly related to resolving ambiguity in meaning (Happé & Frith 2006, Happe 1997, Frith & Snowling 1983, O'Connor & Klein 2004, Martos et al. 2013), identifying pronoun referents (O'Connor & Klein 2004), figurative-language comprehension (MacKay & Shaw 2004), making pragmatic inferences (Norbury 2014), as well as lexical (Speirs et al. 2011) and syntactic (Whyte et al. 2014, Martos et al. 2013) processing.

The next section presents related work from the field of readability assessment.

2.3 Readability

This section presents related work in the field of readability research, including various definitions of the term readability, classic formulae, novel ways of assessing readability based on machine-learning algorithms and finally, the cognitive paradigm in readability research and readability assessment for people with developmental disorders.

2.3.1 Defining Readability

As many other constructs, readability does not have a single, universally accepted definition. Lorge (1944) defines the criterion for readability as "the success that large numbers of persons have in comprehending the text", where reading comprehension is the interaction between readability and reading ability. Reading ability per se is defined as "a person's success with an adequate reading test" (Lorge 1944). In Lorge's explanation, the emphasis is on the statistical nature of readability: it is certified by the success of large numbers of people.

In terms of explaining the criterion and premises of readability, some definitions focus on the causal relationship between comprehension and writing: "the ease of comprehension because of style of writing" (Fry 2004). Other

definitions describe readability through the purpose of the assessment: "The purpose of readability assessment is to effect a 'best match' between intended readers and texts; thus, optimal difficulty comes from an interaction among the text, the reader, and his/her purpose for reading" (Chall & Dale 1995). This definition includes the very important premise of the "specific purpose" of the reading process, while simultaneously emphasising the interactivity of the readability construct.

Finally, a definition that embraces 1) the reading ability of a particular reader, 2) the multidimensional nature of the construct of readability, and 3) the purpose of the act of reading, is proposed by Pikulski (2002):

> "*A more reasonable definition of readability is the level of ease or difficulty with which text material can be understood by a particular reader who is reading that text for a specific purpose. Readability is dependent upon many characteristics of a text and many characteristics of readers. Thus, one important characteristic of a useful, informed definition of readability is that it reflects the interactive nature of the construct.*"

In this thesis we adopt this definition by Pikulski (2002), as it takes into consideration the multiple dimensions of the construct of readability.

In the earliest stage of readability research, readability was predicted by specially calibrated formulae. A readability formula is "simply a mathematical equation derived by regression analysis" (McLaughlin 1969). The variables in this equation are the difficulties experienced by people reading a particular text and the linguistic characteristics of this text. Since the formulae are derived from the profile of a reading material and the profile

of a reader population reading this material, there are many formulae designed to capture the characteristics of specific genres and registers as well as of groups of readers - from schoolchildren to adults and from teachers to military clerks. Formulae are validated on norming passages of texts with a known level of difficulty for a particular reader population. A formula itself uses as variables linguistic features with the highest possible discriminative power, i.e. features which can reliably tell a hard text from an easy one and point out as many nuances that differentiate them. The created formula is a statistical tool that can predict the relative difficulty that a similar reader population would experience while reading a similar text. The result is then reported as a numerical index representing either a grade level (referring to years of schooling as a criterion to match readers to text) or a difficulty level of the text.

With the different resources available over time and the evolution of readability research itself, three trends naturally occurred in time. Benjamin (2011) divides them into **traditional methods, methods inspired by cognitive science** and **methods based on the use of statistical language-modelling tools**. The so called "traditional" or "classic" methods refer to the readability formulae that are traditionally concerned with capturing only the quantitative characteristics of a text, such as number of words and sentences. The cognitive methods are motivated by the idea that reading comprehension is determined by more sophisticated factors such as propositions and inferences (Kintsch & Vipond 1977). Finally, the methods

based on statistical language modelling emerged thanks to the development of artificial intelligence.

2.3.2 Early Readability Formulae

The first readability formula was developed by B. A. Lively and S.L. Pressey in 1923 and aimed to assess the readability of textbooks used in schools at that time (Lively & Pressey 1923). This formula became a basis for the development of other early formulae, such as the Winettka formula (Vogel & Washburne 1928), which introduced sentence features, and the Dale and Tyler formula (Dale & Tyler 1934), which was the first to assess reading materials for adults.

The culmination of early forays into readability research to identify the full spectrum of factors that define text complexity was Gray and Leary's study "What makes a book readable" (DuBay 2008). The authors identified 228 elements of reading ease and grouped them into four main categories: *Content* (propositions, organisation, coherence), *Style* (semantic and syntactic elements), *Format* (typography, illustrations), and *Organisation* (chapters, headings, navigation).

Gray and Leary's extensive research was used as a basis for the development of the Lorge Readability Index (Lorge 1944) and the first Dale-Chall formula (Dale & Chall 1948), which both relied on word lists to measure word familiarity. In the case of (Dale & Chall 1948), the word list consisted of 3,000 easy words, familiar to children in Year 4. Examples of easy words

from the Dale and Chall list of 3, 000 easy words include:

E: *each eager eagle ear early earn earth east eastern easy eat eaten edge egg eight eighteen eighth eighty either elbow elder eldest electric electricity elephant eleven elf elm else elsewhere empty end ending enemy engine engineer English enjoy enough enter envelope equal erase eraser errand escape eve even evening ever every everybody everyday everyone everything everywhere evil exact except exchange excited exciting excuse exit expect explain extra eye eyebrow*

K: *keen keep kept kettle key kick kid kill killed kind kindly kindness king kingdom kiss kitchen kite kitten kitty knee kneel knew knife knit knives knob knock knot know known*

Q: *quack quart quarter queen queer question quick quickly quiet quilt quit quite*

The formula is then computed as follows:

$$RawScore = 0.1579 \times (PDW) + 0.0496 \times (ASL) + 3.6365 \tag{2.1}$$

Raw score in the formula stands for the uncorrected reading grade[1] of a student who can answer half of the test questions on a passage, PDW stands for "Percentage of difficult words not on the DaleChall word list" and ASL stands for "Average sentence length".

Despite the fact that, for a long time after it was created, the original Dale

[1]US class grade

and Chall formula (1948) was considered one of the most precise measures of text difficulty, in 1995 the authors published a revised version of it (Chall & Dale 1995).

$$RawScore = 64 - (.95 \times NUW) - (.69 \times ASL) \tag{2.2}$$

The two basic features of the original formula (Number of Unfamiliar Words (NUW) and Average Sentence Length (ASL)), were preserved. The revision of the original formula was made in two main directions: 1) the validation of the formula against new, more precise sets of criterion passages and an updated word list and 2) the simplification of the way the formula is computed. Among the tools used for the validation were various criterion passages, notably the Bormuth passages (Bormuth 1971). The Dale-Chall formula was cross-validated with many other tests and formulae including the Fry graph[2] (Fry 1968) and the average judgments of teachers on 50 passages of literature. The reading levels varied from 1 to 16, with 1 corresponding to first grade and 16 to college graduate level. The revised Dale-Chall formula and the Bormuth Mean Cloze Score have a correlation of .92 (Bormuth 1971).

Another formula from the early period is the Flesch Reading Ease formula (Flesch 1948), which also had an early version, from 1943, and a revised version, from 1948. The latter formula had two stages: the first (part A) assesses the reading ease of the text by counting the average sentence length

[2]A method for measuring readability using a graph, developed by Edward Fry in 1963, 1968

in words and average word length in syllables:

$$FRE = 206.835 - 1.015 \times \frac{words}{sentences} - 84.6 \times \frac{syllables}{words} \qquad (2.3)$$

The rationale for this, as in the initial formula, was that "measurement of word length is indirectly a measurement of word complexity and that word complexity in turn is indirectly a measurement of abstraction" (Flesch 1948). The second stage, (part B) aimed to measure human interest by counting the number of personal words (pronouns and names), as well as the number of personal sentences (quotes, exclamations, incomplete sentences). Flesch (1948) replaced the count of affixes with the average number of syllables per word, which, in his own words, were "obviously easier to count than affixes since this work can be reduced to a mechanical routine" (Flesch 1948). The deliberate separation of the new formula in two parts aimed to distinguish how the two factors, isolated from one another, correlated and contributed to the measurement of readability. While part A had a correlation coefficient of .74, part B scored much lower, with a correlation coefficient of only .43. Flesch (1948) stated that his rationale for including this element: "the correlation coefficient shows only to what extent human interest in a given text will make the reader understand it better." (Flesch 1948). The overall value of the formula was not only its high correlation coefficient (Part A), but also that through the variable "human interest' (Part B) it for the first time took into account features such as the reader's attention and motivation for continued reading.

Another formula from the early period is the Fog Index (Gunning 1952). The name "Fog" was used as a metaphor for the unnecessary complexity of texts in newspapers and magazines (DuBay 2008). Like most researchers before him, Gunning used average sentence length as a variable.

$$GL = 0.4 \times (ASL + HW) \tag{2.4}$$

In this equation, the Grade Level (GL) is assigned through average sentence length (ASL) and number of hard words (HW) for each 100 words of a document. What makes the computing of this formula simpler in comparison to previous formulae is that, instead of word lists, the formula utilises the number of "hard words", which are words with more than two syllables.

Along the path of creating refined readability formulae that explore new dependencies between the well-known "golden" elements of word and sentence length, Edward Fry (1963, 1968) created a new method to test readability using a graph (Fry 1968). This is probably the most easy-to-use readability measure, as the only thing that the assessor is required to do is to count the average number of sentences and words in a 100-word sample and to place the number on the vertical and horizontal axes of a graph.

Another tool which follows the trend of "simplicity and ease' is the humourously titled "Simple Measure of Gobbledygook" or SMOG Grading formula (McLaughlin 1969).

$$SMOG = 3 + \sqrt{PolysyllableCount \times \frac{30}{SentenceNumber}} + 3.1291 \quad (2.5)$$

The name of the formula is also a reference to the Fog index (Gunning 1952), which was the first to count "hard words" as words with more than 3 syllables, and his index is used as a basis for the SMOG formula. What differentiates the two formulae is that in SMOG these variables are multiplied instead of added. The argument for multiplying them is the nature of the interaction between semantic and syntactic difficulty: "A slight difference in word or sentence length between two passages does not indicate the same degree of difference in difficulty for hard passages, as it does for easy passages." (McLaughlin 1969).

It was during the 1970's that the U.S. Army started funding extensive readability research on technical manuals, administrative documents and scientific literature. The first formula produced as a result of this research was the FORCAST formula, (Caylor et al. 1973), which was especially designed for the Army and is easy for standard clerical personnel to apply without special training (DuBay 2008):

$$GradeLevel = 20 - (\frac{N}{10}) \quad (2.6)$$

In this equation, N stands for the Number of Single-syllable Words in a 150-word Sample.

What is interesting about it is that it is the first formula not to include

sentence length as a readability factor but to use only the number of one-syllable words as such. Nevertheless, it correlates highly with both the Flesch formula (Flesch 1948) and the Dale-Chall formula (Dale & Chall 1948) and proved that there was a literacy gap between the level of complexity of Army reading materials and the reading ability of the Army personnel at that time. Owing to the fact that it does not include sentence length as a factor, the FORCAST formula is suitable for assessing short statements, applications and forms.

The FORCAST formula was the first of a whole sequence of Army-funded readability indices, among which are The Army's Automated Readability Index - ARI (Senter & Smith 1967), The Navy Readability Indices, also known as the Flesch-Kincaid readability formula (Kincaid et al. 1975), the Hull Formula for technical writing (Hull 1979), etc. The Hull formula (Hull 1979) was developed owing to the observation that technical literature consists of a lot of "hard words" and terms and that, therefore, to predict its readability, a new formula, that does not rely on word length, was needed. While sentence length was confirmed as a valid predictor for the readability of technical texts, "Hull claims that an increase in the number of adjectives and adverbs before a noun lowers comprehension. His study indicates that the modifier load is almost as predictive as a syllable count, more causal, and more helpful for rewriting" (DuBay 2008).

Another formula developed for the Army, the Flesch-Kincaid formula (Kincaid et al. 1975), was actually a recalculation of ARI, Flesch and the

Fog index, customised for the requirements of the Navy. The new indices traditionally use sentence length and word length in syllables:

$$FKGL = 0.39 \times \frac{words}{sentences} + 11.8 \times \frac{syllables}{words} - 15.59 \qquad (2.7)$$

The original contribution of this study, apart from the high accuracy of the recalculated indices, is that it investigated another factor for predicting readability: learning time. It discovered the correlation between the time students need to learn information from a text, their pre-assessed reading ability and the level of complexity of the text. As a result, it was confirmed that not only reading ability but also learning time can reliably predict readability. Moreover, by using both comprehension scores and learning times, the Flesch-Kincaid formula was able to differentiate significantly between texts, that are less than one grade level apart. Nowadays the Flesch-Kincaid formula is broadly used as a readability measure for all kinds of documents.

2.3.2.1 Limitations of the formulae

Without a doubt, readability formulae had a significant influence on research and on the publishing business, and to a great extent changed the way people think about communication through written text. However, as popular as they are, these formulae have a number of limitations that question their reliability as a metric. First and foremost, they are criticised for employing only superficial characteristics of text such as word and sentence length. By proposing a correlation between word length and word complexity, readabil-

ity formulae cannot determine the complexity of ideas and the logical order of the content (Bruce et al. 1981, Gilliland 1972, Kintsch & Vipond 1977).

Subtle characteristics of the text like cultural bias, style, implicitness of ideas and whether the text is stimulating for the reader or not, also remain out of range for the readability formulae. While these are features of the text, the formulae also fail to measure variables related to the reader or his or her prior knowledge, level of interest and the purpose and circumstances in which the text is being read.

Another problem related to readability formulae is their misuse (Siddharthan 2004, Benjamin 2011). An important moment in the "for and against" debate is that one has to know when and how to apply them. Siddharthan (2004) notes that many teachers over-trusted the formulae, which resulted in their students' being denied access to "too-complex" materials, despite the students' interest in reading them. Moreover, the formulae were widely applied to genres like poetry where they could not show, and were not meant to show, accurate prediction due to the different linguistic structure and purpose of reading. What Siddharthan (2004) rates as the most damaging misuse of the formulae is that instead of applying them post hoc, many authors used them to adjust their writing to the intended audience:

> "*By the 1970's, a situation had been reached where the use of readability formulae at the authoring stage had resulted in dozens of unreadable textbooks. The problems arose because authors were subconsciously manipulating sentence and word lengths without decreasing the syntactic or lexical complexity*"

There are many other types of misuse of the formulae apart from using

them ad hoc or taking them as an absolute measure. One such misuse is the application of the wrong formula. In many cases, readability formulae are developed with certain purposes in mind and validated on certain types of documents (e.g. the Navy formulae were validated on documents relevant to the Navy). It is questionable to what extent applying a highly specialised formula to a different genre or domain would yield reliable results.

One possible reason for the drawbacks of readability formulae is the very idea that the text-reader interaction could be captured throughout a simple equation. The formulae also rely on the assumption that whatever the output is, it has to be some precise measure, a set of categories in which readership could be neatly divided into. This assumption, called "Aura of precision" (Duffelmayer, 1985 in (Pikulski 2002)), deprives the construct of readability of its most essential characteristics outlined in the beginning: its nature as an interactive process, the consideration of the individual reader and his or her purpose of reading. However, as Pikulski observes, "it seems necessary to conclude that, to date, no objective, accurate way of measuring the concept of readability has been devised." (Pikulski 2002).

Nonetheless, the question of what makes a text readable and how people read is still intriguing for researchers from various fields.

In the following section we discuss machine-learning approaches to readability research and compare them to the traditional methods presented so far.

2.3.3 Machine-learning Approaches in Readability Research

Some of the important innovations that statistical language modelling introduces are new features, enabled by different Natural Language Processing techniques (e.g., average parse-tree height, average distance between pronouns and their anaphors, etc.) as well as new algorithms and machine-learning techniques. Readability assessment becomes a part of various evaluation techniques for text summarisation (Wan et al. 2010), information retrieval (Kane et al. 2006, Yan et al. 2006, Newbold & Gillam 2010), text simplification (Štajner, Mitkov & Pastor 2014, Aluisio et al. 2010, Dell'Orletta et al. 2011), and machine translation (Stymne et al. 2013).

Due to the vast amount of existing literature, the current chapter is limited to areas relevant to the overall aim of the thesis: readability metrics for people with autism. First, we introduce the main approaches from the field of Natural Language Processing (NLP) and then specifically discuss Statistical Language Models and Support Vector Machines, applied to readability research. Then, we focus on the assessment of web documents and the connection between text simplification (TS) and readability assessment at both document and sentence level. This connection is dual: not only readability assessment is used as an evaluation metric for text-simplification tools, but also, readers with cognitive disabilities, who are the target group of the metric proposed in this work, are very often the target group of text simplifi-

cation. We also discuss readability assessment for bilingual education, owing to the fact that many of the features developed for second-language learners, addressing grammar and syntax are also relevant to people with autism.

Another criterion for the works described in the current section is that, with a few exceptions, they explore readability assessment exclusively for the English language. As readability metrics for people with disabilities are of the utmost relevance to the current thesis, they are specifically discussed in Section 2.3.4.

2.3.3.1 Assessing readability with statistical-language models and support vector machines

Natural Language Processing (NLP) is an interdisciplinary field between computer science and linguistics, concerned with the way natural languages are processed by computers. Main approaches in NLP are rule-based approaches, machine-learning approaches or hybrids between the two. Rule-based approaches are large sets of manually-encoded rules given to a machine by a human, while machine-learning approaches make use of comparatively large amounts of data to "learn" rules automatically on the basis of processing many manually annotated labels. Different machine-learning algorithms include decision trees (Quinlan 1986), Statistical Language Models (SLM) (Rosenfeld 2000) and Support Vector Machines (SVM) (Joachims 1998), among others.

Statistical Language Modelling can be defined as "the attempt to capture

regularities of natural language for the purpose of improving the performance of various natural language applications" (Rosenfeld 2000). A statistical language model is a probability distribution P(s) over strings S, which exploits patterns of use in language and aims to reflect how frequently a string S occurs as an entity (word, phrase, sentence, etc.).

An important prerequisite for building a language model is the existence of a large sample of examples of the investigated phenomena. If such data are available, the next step is for them to be statistically analysed so that labels can be assigned to different linguistic categories. For the task of readability assessment, a language model is built for each reading level (Heilman et al. 2008). That is, it is held that each reading level has a set of words or combination of words (n-grams) which typically occur in it. By having this information, the LM can predict the probability that a certain word or combination of words will appear in this reading level. LMs have several advantages over readability formulae. Heilman et al. (2008) argue that language modelling "provides probability distribution across all grade levels, not just a single prediction" and that it can also supply "more data on the relative difficulty of each word in a document" (Heilman et al. 2008). In the next subsections we discuss readability metrics based on statistical language-modelling and the fields in which they are most successfully applied.

One of the widely used algorithms in both general NLP and Readability Machine Learning (ML) is Support Vector Machines (SVM). One application of SVMs is their use as a supervised approach for the classification of

texts (Joachims 1998). Based on a set of labelled training examples, an SVM "learns" how to assign the right label for an unseen document. In the case of readability assessment, training examples would involve sets of features extracted from passages or texts for which the reading level is known; the categories would be the reading levels. In SVMs each training example is represented as a point in an n-dimensional space with a particular position. The task (in linear classifiers) is to divide these points by finding the hyperplane which separates the largest margin between the two closest positions in the space, called support vectors. After the training is over, the system is given a new set (test set), in which labels pointing to the true category of a text are removed. Then the task of the system is to assign the correct categories "learned" during the training process to the new unlabeled examples. SVMs perform testing of different combinations of features in order to find the best combination, meaning that they would exploit the features determining the optimal hyperplane in the graphical representation.

The next sections discuss applications of language modelling and SVM to different areas of readability research.

2.3.3.2 Readability of web content

Language models assign probabilities to the observed frequency of occurrence of the token sequences. Depending on the number of tokens in a sequence, the models can be unigram and n-gram (bigram, trigram, etc.). By considering only one token, unigram models assume that the probability of generating a

word is independent of its context, that is, words which come before or after it. Despite unigram models being weak representations (because in reality, the probability of a word does depend on the context), they are shown to be particularly relevant in the domain of web documents.

One problem with assessing web pages is that they are often too short (it should be noted that readability formulae normally take passages longer than 100 words) and contain a lot of noise such as navigation words from menus, hyperlinks, e-mail addresses, copyright descriptions, etc. These stylistic peculiarities may cause traditional readability formulae to produce an inaccurate assessment of the readability of web text. (Gottron & Martin 2009, Kanungo & Orr 2009). Si & Callan (2001) note that:

- There is a demand for an assessment of web documents free of the bias of the formulae
- Such assessment should account for content (as opposed to surface features like word and sentence length only)
- The assessment tool (in this case the unigram model) should be derived from actual corpora

To satisfy these prerequisites and to test the reliability of sentence and word length as variables in readability assessment in the domain of educational web pages, Si and Callan (2001) compiled a small corpus of 91 web documents. They calculated the Sentence Length and Word Length in Syllables (SLS and WLS) distribution for three readability levels in the corpus.

The results showed that mean values of sentence length increased monotonically throughout the three reading levels while mean values of word length did not: "web pages written for grades 3-5 had more polysyllable words than web pages written for grades 6-8" (Si & Callan 2001). This study showed that one of the most reliable features in the classic formulae, namely Word Length in Syllables, is not as reliable as it had been thought to be, or, in the best case, not as suitable for the domain of web texts as it might have been for other domains. Future readability research, which will most likely explore modern media such as web pages and forums, should use WLS with caution (as well as formulae that have WLS as their main variable (e.g. SMOG formula (McLaughlin 1969)).

The Smoothed Unigram Model by Collins-Thompson & Callan (2004, 2005) categorised individual texts by comparing them to language models of different reading levels. An application of this model was the provision of young students with search tools that can find documents relevant both to their search query and their reading level.

The syntactic component in web documents is not reliable owing to the particular characteristics of web texts (hyperlinks, navigation menus, e-mail addresses, etc.), which is why the model of (Collins-Thompson & Callan 2005) only aimed at assessing the semantic (lexical) component of web texts.

Thus, relying exclusively on the semantic component, which is believed to have sufficient discriminatory power between grade levels, the authors built 12 language models corresponding to the 12 American grade levels. These

models attempted to capture:

1) Tokens or any word occurrence in the corpus

2) Types, which are word strings or linguistic patterns, each of which is counted only once regardless of the number of occurrences.

Their Smoothed Unigram measure shows a good correlation of .67 with grade level and was shown to perform better than a number of other measures, one of which was again the Flesch - Kincaid formula (Kincaid et al. 1975) (similar to (Si & Callan 2001)). Another advantage of the classifier, which makes it applicable to other domains, is that it could be trained on different collections of data.

2.3.3.3 Readability for second-language learners

Foreign-language learning (L2 learning) has a number of characteristics that make it a challenge for readability research (Bennöhr 2005, Ozasa et al. 2007, 2008, François & Fairon 2012). As we shall see below, such specifications are the simultaneous acquisition of grammar and vocabulary (Callan & Eskenazi 2007) and the necessity of a high interest level (Schwarm & Ostendorf 2005). Callan & Eskenazi (2007) suggest that a good improvement in readability assessment for L2 learners is the involvement of pedagogically motivated grammatical features: Passive voice, Past participle, Perfect tense, Relative clause, Continuous tense, and Modal.

The rationale behind this is the observation that in L2 education grammatical rules and vocabulary are acquired simultaneously, while in first lan-

guage (L1) most grammar is mastered in childhood and thus, unlike in L2, complex grammatical constructions are seen in texts with both low and high reading levels. As a result of this, the syntactic components are crucial for assessing L2 learning materials. For this reason, language models for bilingual education are n-gram models, which capture more complex linguistic phenomena including syntactic constructions, but unlike unigram models, are more prone to be affected by semantic noise and scarcity of data.

Callan & Eskenazi (2007) show that a language modelling approach alone was more accurate than a grammar-based approach alone. A possible explanation for this result is that the grammar-based prediction is more negatively affected by the noise in the corpora. In the unigram language model noise would affect only separate words, while in the grammar-based approach (relying on accurately parsed dependencies between words) it may affect a whole clause or even a sentence. Second, a significant advantage of the unigram model is that it relies on single words as predictors and thus can utilise all appearing words as features. The grammar-based approach is more vulnerable in this sense as its prediction components are a finite set of manually chosen features. However, Callan & Eskenazi (2007) also find that grammatical features play a more significant role in readability measures for L2 than for L1.

Schwarm & Ostendorf (2005) propose a method for addressing a specific need of teachers in the domain of bilingual education: "bilingual education instructors seek out "high interest level" texts at low reading levels, e.g.

texts at a first or second grade reading level that support the fifth grade science curriculum" (Schwarm & Ostendorf 2005). The authors find that methods based on word lists (such as the Dale-Chall formula (Chall & Dale 1995)) are not suitable for this task because the appropriate reading material should contain relatively difficult and topic-specific words, while the syntactic and discourse structure should be kept simple. In other words, Schwarm & Ostendorf (2005) hypothesise that for the purpose of enhancing students' vocabulary, there is a discrepancy between the reading level of the lexical and the syntactic components of the materials. Language models, like the one proposed by Collins-Thompson & Callan (2005) are also not applicable to this task because they rely solely on the lexical component.

The method of Schwarm & Ostendorf (2005) relies on Support Vector Machines, combining traditional features with features from parse trees and statistical-language models. In their experiments, no feature was found to be more important than any other and, moreover, "performance was degraded when any particular features were removed" (Schwarm & Ostendorf 2005).

Petersen & Ostendorf (2007) expand L2 assessment research towards generalising of the Schwarm & Ostendorf (2005) classifier to handle new data and towards continuing to investigate the relationship between syntactic and lexical features for second-language acquisition. In L2 acquisition "even intermediate and advanced students of second languages, who correspond to higher L2 reading levels, often struggle with the grammatical structures of their target language" (Callan & Eskenazi 2007). The advantage of this

distinction between L1 and L2 readers is that the grammatical features are established to be a component which introduces significantly higher complexity for L2 readability assessment. It is important to note, however, that vocabulary also plays a crucial role for L2 learners. The distinction made is rather that grammatical features are more relevant to readability assessment for L2 than L1 learners. This distinction is also relevant to readers with autism, who have been shown to to find both lexico-semantic and syntactic components of texts challenging (Section 2.2.2).

The results shown by Petersen & Ostendorf (2007) suggest that new unlabelled data can be used to augment the corpus but that inclusion of syntactic features in the SVM has a relatively small effect on the overall performance. Unlike Petersen & Ostendorf (2007), François & Fairon (2012), who develop a readability formula for French as a foreign language, assume that not syntactic features but semantic ones cause redundancy. Like other scholars from the early research period, they come to the conclusion that "maximizing the type of linguistic information might not be the best path to go" (François & Fairon 2012). In their experiments, a simple model of four variables outperforms a more elaborate one.

The next section explores the role of readability assessment as an evaluation technique for text simplification.

2.3.3.4 Sentence-level readability assessment and evaluation of text simplification

Text simplification (TS) is a process which aims to enhance the understandability of a text by performing different linguistic transformations without changing the original meaning of the text (Max 2000). It is particularly useful for people with disabilities as evidenced by the number of TS projects for people with various conditions such as:

- Autism (FIRST project [3])
- Dyslexia (DysWebxia project [4])
- Down syndrome (Simplext project [5])
- Mild cognitive impairment (READ IT (Dell'Orletta et al. 2011))

In recent years, automatic simplification tools are gaining more and more popularity (Dell'Orletta et al. 2011, Inui et al. 2001, Aluisio et al. 2010). A number of publications discuss the place of readability research in the development and evaluation of such tools (Štajner, Mitkov & Pastor 2014, Dell'Orletta et al. 2011, Petersen & Ostendorf 2007, Aluisio et al. 2010). Readability metrics are suitable for the evaluation of the output text both by comparing it to the source text and by measuring the usefulness of the simplified version to the target reader population.

[3]FIRST project. Available at: http://www.first-asd.eu/

[4]DysWebxia project. Available at: http://www.luzrello.com/DysWebxia.html

[5]Simplext project. Available at: http://www.simplext.es/

Aluisio et al. (2010) propose an approach towards readability assessment for text simplification for Portuguese, the aim of which is to distinguish between original and simplified texts, as well as between two different levels of simplification: natural (including only slight modifications) and strong (including more elaborate modifications). The feature set used for the development of the tool consists of cognitively-motivated features from Coh-Metrix PORT (a version of Coh-Metrix for Brazilian Portuguese), as well as of syntactic features and n-gram model features. As the main function of the tool is to assess the readability of automatically simplified texts, some of the syntactic features included are specially designed "to capture simplification operations" (Aluisio et al. 2010). These reflect the incidence of clauses, adverbial phrases, appositions, the passive voice, relative clauses, coordination, and subordination. The results show that the tool distinguishes more successfully between original and simplified documents than between the two types of "natural" and "strong" simplification.

Dell'Orletta et al. (2011) also discuss readability assessment with respect to text simplification. The specifications of their approach are in accordance with a particular application of their system: it is aimed at the Italian language and readers with low literacy skills or Mild Cognitive Impairment (MCI). As in previous research, the authors treat readability assessment as a binary classification task aiming to distinguish between easy-to-read and difficult-to-read documents.

While Aluisio et al. (2010) work with whole documents and include spe-

cific syntactic features to tackle the transformation from original to simplified text, Dell'Orletta et al. (2011) approach the same task by carrying out assessment on two linguistic levels: sentences and documents. Trying to assess the readability of particular sentences imposes various difficulties, such as the lack of suitable sentence-level measures, the lack of training data and last but not least, the debatable relationship between sentence length and discourse features with regard to text complexity.

To overcome these limitations, the authors perform an experiment where "READ-IT", their SVM-based classifier, has to detect easy-to-read sentences within a difficult-to-read document. This type of evaluation is based on the notion of Euclidean distance between vectors, where each feature has a vector representing a set of sentences; and the smaller the distance between vectors is, the more similar the sets of sentences are. READ-IT shows high accuracy in document classification. Among the four models involved in the classifier (base, lexical, morpho-syntactic and syntactic) the morpho-syntactic one shows highest accuracy for document classification (98.12%). At the level of sentence classification the accuracy is much lower, though still encouraging with the highest performance achieved by the syntactic model (78.2%).

Other studies that report on sentence-level readability assessment include a project for deaf students, where a readability model was built to classify pairs of original and manually simplified sentences using training examples classified by teachers (Inui et al. 2001). The model described in this paper

achieved 95% precision and 89% recall (Inui et al. 2001); however, this very high accuracy may be due to the fact that the simple sentences were not naturally occurring but manually simplified by following a specific set of instructions for reducing syntactic and lexical complexity, which may have been easily captured by some of the features of the model. Other attempts to classify sentences based on readability achieved 80% accuracy by using pairs of original and manually simplified sentences from news articles (Vajjala & Meurers 2014) and 71% accuracy when classifying Swedish sentences for foreign language-learners (Pilán et al. 2014).

The next section discusses approaches from the cognitive paradigm in readability research and aspects of readability related to people with developmental disorders.

2.3.4 Addressing Reader-related Aspects of Readability: Cognitively-based Analysis and Readers with Disabilities

Classic readability formulae and machine-learning approaches do not account for the cognitive processes which underlie reading. Both classic and machine-learning approaches, with a few minor exceptions, rely on graded passages and try to find those combinations of text properties that would best match this grading. The drawbacks of using such superficial features include the inability of the formulae to measure the prior knowledge of the readers, how well

the ideas in the text are organised and what cognitive load the text imposes on the reader. The following section will focus on exploring the cognitively-motivated features in readability and how the psychological characteristics of particular user populations are taken into consideration for the development of appropriate metrics.

In this section, we discuss various techniques for measuring coherence and reading ability as an individual characteristic. First, we present the concepts of propositions and inferences (Kintsch & van Dijk 1978) as properties of the text, which contribute to understanding the amount of effort required by the reader to comprehend a piece of text. We present novel approaches to the matching of readers to texts, namely Latent Semantic Analysis (Landauer et al. 1998) and Coh-Metrix (Graesser et al. 2004, McNamara et al. 2014). We present cognitively-motivated features included in the MRC database (Coltheart 1981), as well as cognitively-motivated features addressing specific profiles of readers with disabilities, namely those with Intellectual Disability (Feng 2009, Feng et al. 2009, 2010) and dyslexia (Rello, Baeza-Yates, Bott & Saggion 2013, Rello et al. 2012, Rello, Baeza-Yates, Dempere-Marco & Saggion 2013). At the end of this section we present related work on readability assessment for people with autism (Štajner, Mitkov & Pastor 2014).

2.3.4.1 Propositions and inferences

To capture more accurately the role of the psychological construct of coherence (how chunks of text and their meaning are connected in the reader's

mental representation), research from the structuro-cognitivist paradigm aims to measure cohesion: a text property referring to the surface indicators of how sentences are related to one another in a text (Benjamin 2011).

Kintsch & van Dijk (1978) argue that "the semantic structure of texts can be described both at the local microlevel and at a more global macrolevel". The smallest units in these levels, that can carry meaning, are propositions: units of a predicate and at least one argument. Propositions can contain other propositions and can refer to one another. The extent to which a text is cohesive is measured by the overlap of propositions (or at least their arguments) in consecutive sentences in the microlevel and throughout a larger part of the text in the macrolevel. Thus, if there are many explicitly overlapping propositions, there is less effort required by the user to link them and to follow the discourse, which is recommended for texts for novice readers.

McNamara & Kintsch (1996) apply this theory in experimental conditions and come to the conclusion that readers with lower prior knowledge of a certain topic would benefit from reading a high-cohesion text, while more experienced readers would benefit from a low-cohesion text, which would require them to make more inferences. Moreover, the study provides scientific evidence for the fact that comprehension may suffer not only if readers are presented with a text that is too difficult but also, if they are presented with materials below their reading level.

Britton & Gülgöz (1991) use the Kintsch & van Dijk (1978) model to modify texts by locating parts of the text where propositions do not overlap.

The changes they make to the texts are as follows:

- link each sentence to the previous one via overlapping propositions
- fill these gaps by using only one term for each concept
- arrange the sentences so that old information precedes new information
- make implicit inferences explicit

Crucially, these changes were not captured by any readability formulae applied to the original and modified texts: the Flesch-Kincaid formula (Kincaid et al. 1975) and the ARI formula (Senter & Smith 1967), for example, rated the two versions the same. Participants, however, demonstrated much better free recall on the modified texts and got a higher comprehension score on the multiple-choice questions.

2.3.4.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) (Landauer et al. 1998) is a method for extracting the meaning of words as they appear in context and representing them as vectors in semantic space. Much like statistical approaches, LSA uses large corpora to learn the likelihood of certain words appearing in a particular context. Each word is represented as a vector in the semantic space, with rows in the vector representing different contexts. The cosine between two vectors provides a numerical value of their relationship (Benjamin 2011).

Unlike other statistical approaches, LSA provides an opportunity to match readers to texts on the basis of their prior knowledge. This is achieved

by measuring the semantic relatedness between a short text written by the reader and other texts on the same topic, which might be of interest to them.

Research based on LSA shows that "texts in which there is a high degree of cohesion tend to be easier for non-expert readers to read than texts in which more connections have to be made by the reader" (Benjamin 2011). Unfortunately, there are still some drawbacks in the LSA theory such as lack of representation of the human cognitive abilities used to construct and apply knowledge, and "the ability to use detailed and complex order information such as that expressed by syntax and used in logic" (Landauer et al. 1998).

2.3.4.3 Coh-Metrix

Coh-Metrix is a tool developed by a research team in the Department of Psychology, University of Memphis (Graesser et al. 2004, McNamara et al. 2014). It includes a large span of cognitively-based features. The latest version, Coh-Metrix 3.0[6], contains 108 indices grouped into the following 11 categories:

1. Descriptive features
2. Text easability principal component scores
3. Referential cohesion
4. LSA features

[6]Coh-Metrix 3.0 (tool). available online at: http://tool.cohmetrix.com/

5. Lexical diversity
6. Connectives
7. Situation model
8. Syntactic complexity
9. Syntactic pattern density
10. Word information
11. Readability formulae

The descriptive measures of readability include the number of sentences and the number of paragraphs, as well as measures of different types of cohesion (referential, deep, verb cohesion, etc.), measures of connectivity (number of explicitly conveyed logical relations) and temporality (number of cues about temporality). Latent Semantic Analysis, described above, is also included in Coh-Metrix 3.0. Lexical diversity, measured via the type-token ratio gives information about the variety of unique words that appear in text in relation to the total number of words.

Except for traditional measures of syntactic complexity, Coh-Metrix also accounts for the density of particular syntactic patterns and includes readability formulae such as the Flesch Reading Ease formula (Flesch 1948) and the Flesch-Kincaid readability formula (Kincaid et al. 1975). The word-information index refers to variables such as age of acquisition, familiarity,

concreteness, imagability and meaningfulness of words, the measurement of which is discussed in the next subsection.

2.3.4.4 The MRC Psycholinguistic Database

The Medical Research Council (MRC) Psycholinguistic Database (Coltheart 1981) is a computerised system containing cognitive measures for a total of 98,538 words. These measures are obtained by presenting stimulus words (without context) to large numbers of subjects (usually college students) who are then asked to evaluate word properties such as how easy or difficult it is mentally to picture the word referent (imagability), how abstract or concrete the word is, the age at which the word is first acquired (evaluated on children), etc. The MRC database also includes norms of word meaningfulness, which relate to the number of meanings a word has, namely the Colorado Meaning Norms (Nickerson & Cartwright 1984) and the Paivio Norms (Paivio et al. 1968). The Colorado Norms are an expanded version of the Paivio Norms and were obtained from 197 college students who were asked to write down as many different meanings of a word as they could think of in 30 seconds. An important distinction between meaningfulness and polysemy is that meaningfulness represents the ease with which human subjects could quickly access various meanings of a word; however, it is important to note that these measures do not take into account the role played by context.

The next sections discuss readability assessment for readers with devel-

opmental disorders, namely intellectual disability, dyslexia and autism.

2.3.4.5 Readability assessment for readers with intellectual disability

Intellectual Disability (ID) or Intellectual Developmental Disorder is a developmental disorder, characterised by deficits in general mental abilities, and by an impairment in adaptive functioning for the individual's age and sociocultural background, where all symptoms must have an onset during the developmental period (American Psychiatric Association 2013). Owing to these cognitive deficits, people with ID experience a range of difficulties related to reading (in cases of mild intellectual disability only, as individuals in the moderate and severe end of the spectrum are usually not able to read).

The most eminent work on readability assessment for people with Intellectual Disability is authored by (Feng 2009, Feng et al. 2009, 2010). These studies point out the need for a readability-assessment tool from the user's perspective and as a text-simplification aid tool. First, from the user's perspective, such a tool is needed because of the lack of appropriate reading materials which are simple enough for the users to understand but at the same time discuss topics relevant to adult life rather than being aimed at children, as is often the case. One of the roles of readability in text simplification is to identify parts of the text that would pose difficulty to the users and, when more than one simplification option is available, to guide decisions on which output would be the most favourable one (Feng 2009). The

technical aspects in the field of Text Simplification are growing at a steady pace, while in many cases the psycholinguistics knowledge of the profile of the target population is lagging behind. In this context, readers with autism are also a target group for text-simplification projects like the FIRST project[7], which is why readability assessment for people with autism has a concrete practical application in the field of text simplification.

In spite of the different aetiology of the disorders, people with ID and people with ASD have a lot in common in terms of reading difficulties. Both groups are slow in resolving the identities of proper names, have difficulty integrating complex information, have limited working memory and are better at decoding words than at comprehending text meaning. Feng (2009) states that "text properties that influence reading difficulty for average readers are qualitatively (but perhaps not quantitatively) the same for readers with ID" and continues, "we are not aware of any text properties that cause problems only for readers with ID" (Feng 2009). The lack of such differences suggests that improved readability metrics for readers with intellectual disability would also improve readability assessment in general. This is yet another example supporting the statement that accessibility for one group of people means improved accessibility for everyone.

To address the ID reading profile, Feng et al. develop discourse-level features extracted from two comparable corpora of paired original documents and their simplified versions for children (LiteracyNet and Encyclopaedia

[7]FIRST project. Available at: http://www.first-asd.eu/

Britannica (Barzilay & Elhadad 2003) and a third corpus from the Weekly Reader (Allen 2009) which is annotated for twelve readability levels). While these corpora are used for the training of the readability tool, the evaluation is performed on a small corpus called LocalNews, specifically developed for the purpose. LocalNews contains twenty articles of news stories simplified by experts and assessed by fourteen adults with ID.

To address better the characteristics of readers with ID, in addition to shallow and syntactic features, the authors include discourse features aiming at measuring the load on the working memory of the reader:

Entity density (person, location, and organisation) addresses the reduced ability of people with ID to keep in mind many entities while reading a text. By entity, the authors refer to the connection between common nouns and named entity noun phrases in the text. It is expected that whole documents, as well as individual sentences with more entities would be harder to encode into semantic memory. The approach counts entities per sentence and per document, which allows not only the assessment of the whole document but also the identification of particular sentences which might pose difficulty.

Lexical Chains indicate synonymy or hyponymy relations between nouns in a text. Lexical chains have "both a length (number of noun phrases it includes) and a span (number of words in the document between the first noun phrase at the beginning of the chain and the last noun phrase that is part of the chain)" (Feng 2009). Lexical chains can also have a state: for a particular word in a text the lexical chain is "active" if the word is in its span.

Overall, the features proposed by Feng (2009) cover relevant issues in the reading ability of people with ID and also show a significant difference between the original and simplified documents. The only two features which score lower are average lexical chain length and number of lexical chains with span greater than half the document. Nevertheless, lexical chains are still proven to be a good predictor of readability, as features like number of lexical chains, average lexical chain span and number of lexical chains active for each word/noun phrase showed a significant difference.

The evaluation of the tool shows that a model based on the novel features only, is outperformed by a model trained on the shallow and parse-related features. However, consistent with results previously presented, a combination of all features performs best.

Training the model on the Weekly Reader corpus (Allen 2009) but testing on the news corpus evaluated by people with ID throws up a surprising result. The optimal model incorporating all features performs worse on the user-evaluated test set than the model consisting of the cognitively-based features only. In the words of the authors, this suggests that "the shallow and parse-related features of texts designed for children are not the best predictors of text readability for adults with ID" (Feng 2009).

2.3.4.6 Readability assessment for readers with dyslexia

Developmental dyslexia is one of the most common reading disorders, characterised by lower reading and academic achievements, which are not caused

by intellectual disability or sensory deficit (American Psychiatric Association 2013). It is considered to be related to non-typical eye fixations, motor control or short-term memory and visual-memory deficits. These deficits lead to reading issues such as a reduced ability to recognise words, and to distinguish between mirror letters like “d” and “b”. they also lead to skipping words in a text and consequently, poor spelling and reading comprehension.

Some interesting factors affecting readability of texts for Spanish readers with dyslexia have been studied by Rello et al. (Rello, Baeza-Yates, Bott & Saggion 2013, Rello et al. 2012, Rello, Baeza-Yates, Dempere-Marco & Saggion 2013). Rello et al. (2012) argue that in the case of dyslexia, readability and understandability should be considered separately. They point out that certain text modifications like the inclusion of graphical schemes, can improve the speed of reading, which they define as an improvement in readability, while others, like frequent or shorter words, may improve reading comprehension, defined as understandability (Rello et al. 2012, Rello, Baeza-Yates, Dempere-Marco & Saggion 2013).

An interesting finding of these studies is that more frequent words improve the readability of the text (readability as defined by the authors), while inclusion of shorter words significantly improves understandability.

Rello et al. (2012) also explore the role of graphical schemes as a device to enhance text readability for readers with dyslexia. A surprising result is the fact that dyslexic and non-dyslexic readers have opposite opinions on the inclusion of graphical schemes. While readers with dyslexia found them

helpful in terms of readability, understandability and recall, non-dyslexic readers from the control group found them misleading and unnecessary. This suggests that, unlike in the case of intellectual disability according to Feng (2009), there is a qualitative difference between the reading aids required for dyslexic and non-dyslexic readers.

2.3.4.7 Readability assessment for readers with autism

While there has been some work on text simplification for people with autism (Evans et al. 2014, Dornescu et al. 2013, Orasan et al. 2013), work on readability assessment for autism is extremely scarce. Owing to the lack of user-evaluated materials, so far the topic has been approached solely as a NLP task without consideration of its psycholinguistic aspects. However, there have been some initial attempts to assess readability for people with autism.

As a part of a larger study on readability for text simplification, Štajner, Mitkov & Pastor (2014) evaluate the discriminative power of the following 18 linguistically motivated features on a corpus of original and manually simplified texts for people with ASD and ID:

1. Average number of verbs per sentence
2. Average number of adjectives per sentence
3. Average number of adverbs per sentence
4. Average number of determiners per sentence

5. Average number of nouns per sentence
6. Average number of prepositions per sentence
7. Average number of coordinating conjunctions per sentence
8. Average number of subordinating conjunctions per sentence
9. Average number of main verbs (verb chains) per sentence
10. Average number of pre-modifiers per sentence
11. Average number of post-modifiers per sentence
12. Average sentence length (measured in words)
13. Average word length (measured in characters)
14. Average number of pronouns per sentence
15. Average number of senses per word
16. Percentage of ambiguous words in the text
17. Average number of senses per word
18. Percentage of ambiguous words in the text

Owing to the lack of user-evaluated materials, the corpus that was used for the evaluation of the applicability of these features to readers with autism was a corpus of 25 original and 25 simplified documents, where the simplification was performed by carers of people with autism. This corpus was developed

as part of the FIRST[8] project. The results of the study show a significant correlation between these indices and the corpus of original and simplified texts used, and support the idea that these features indeed represent reading obstacles for readers with autism.

The next section presents the state-of-the-art in eye-tracking studies for investigating text complexity.

2.3.5 Eye-tracking Methods for Investigating Text Complexity

Eye tracking is a process where an eye-tracking device measures the point of gaze of an eye (gaze fixation) or the motion of an eye (saccade) relative to the head and a computer screen. Fixations are eye movements which stabilise the retina over a stationary object of interest (Duchowski 2009), which, in the case of reading research, is the written text and its units (letters, words, phrases, etc). Gaze fixations and revisits (go-back fixations to a previously fixated object) have been widely used as measures of text processing difficulty by taking into account their durations and the places in text where longer fixations occur (Duchowski 2009).

The idea that the durations of gaze fixations could be used as a proxy for measuring cognitive load dates back to the *strong eye-mind hypothesis* by Just and Carpenter (1980), according to which, "there is no appreciable

[8]FIRST project. Available at: http://www.first-asd.eu/

lag between what is fixated and what is processed" (Just & Carpenter 1980). That is, when a subject looks at something, he/she also processes it cognitively. The hypothesis also states that the amount of time the subject spends on processing the particular object is equal to the amount of time his/her gaze stays fixated on this object.

A series of studies on eye tracking during reading was conducted by Rayner et al. and summarised in (Rayner 1975, 1998, Rayner et al. 2012). The effects of different linguistic constructions investigated included word frequency, verb complexity and lexical ambiguity (Rayner & Duffy 1986), as well as contextual effects on word perception (Ehrlich & Rayner 1981) and the way eye movements reflect attention while reading (Rayner 2009).

Readers have been shown to fixate longer on rare words, words that are semantically ambiguous, and words that are morphologically complex (Rayner et al. 2012). Fixation durations and their number in the text are also affected by other text features such as verb complexity and lexical ambiguity (Rayner & Duffy 1986). These findings are integrated into a model of eye-movement control during reading called the E-Z model (Reichle et al. 1999), which provides a theoretical framework for understanding "how word identification, visual processing, attention, and oculomotor control jointly determine when and where the eyes move during reading" (Reichle et al. 2003).

In terms of corpora containing data from eye fixations, there are only a few relatively large corpora containing eye-tracking data obtained during a reading task. The Dundee corpus (Kennedy et al. 2003, 2013) was developed

as participants read newspaper articles from The Independent or Le Monde newspapers, so the corpus includes whole texts and the languages included are English and French. The Potsdam Sentence Corpus (Kliegl et al. 2004, 2006) is another corpus of eye-tracking data obtained through reading but it focuses on sentence reading only. It comprises records of eye movements from 222 participants reading 144 German sentences. Both these corpora contain eye-movement records from people of average reading ability. To the best of our knowledge, currently there are no corpora available containing eye fixations obtained from people with autism or other disabilities relating to reading. The next subsection presents related work on eye tracking during reading involving clinical populations.

2.3.5.1 Eye Tracking during reading for people with autism and dyslexia

The E-Z model and the majority of eye-tracking research on reading has been conducted on, and is relevant to, the general population of non-impaired readers but has also been applied to clinical populations such as readers with dyslexia (Rayner 1998, Eden et al. 1994) and autism (Sansosti et al. 2013, Brock et al. 2008). Eye-tracking studies involving dyslexic subjects have also been conducted to aid the development of text-simplification systems (Rello, Baeza-Yates, Bott & Saggion 2013), as well as to train machine-learning models to distinguish between dyslexic and non-dyslexic readers on the basis of eye-tracking data (Rello & Ballesteros 2015).

So far, eye tracking has been applied relatively scarcely to the investigation of reading in autism; however, there have been studies investigating whether people with autism process words in context (Brock et al. 2008), where the participants are asked to look at images relevant or irrelevant to a target word while hearing it in a sentence. A study by Sansosti et al. (2013) investigated the ability of autistic adolescents to make bridging inferences. This was done by recording eye-tracking data while the participants were reading pairs of sentences. The data revealed that there was a significant difference between the total fixation durations, number of fixations and number of regressions between the autistic and non-autistic participants (Sansosti et al. 2013).

2.3.6 Summary of Findings

Readability formulae have greatly helped to improve the accessibility of various types of written documents; however, while simple to compute, the formulae can often be inaccurate or misleading. Models based on machine-learning techniques have been shown to be more accurate in their decisions owing to the variety of NLP-enabled features they deploy and their more sophisticated learning algorithms.

As shown earlier in this chapter, the development of both readability formulae and machine-learning readability models depends on the availability of gold-standard data. In the context of readability formulae, these data are the criterion passages, which the formulae are calibrated on. In the context

of machine-learning models, these are datasets used for the training and evaluation of the model. Such datasets did not previously exist for readers with autism. Hence, a first step towards readability assessment for readers with autism would be the collection of a set of texts, the complexity of which would have been evaluated using comprehension testing with people with autism.

Unlike document-level readability assessment, where the gold standard is texts with known difficulty for the target group, the gold-standard criterion for sentence-level readability assessment has not been that clear. Some of the studies in this chapter have used pairs of original and simplified versions of sentences (Inui et al. 2001, Vajjala & Meurers 2014) or unmatched sentences obtained from easy and difficult texts (Dell'Orletta et al. 2011). None of these datasets has been evaluated by the target users, owing to the fact that sentence difficulty could not be evaluated using comprehension questions for individual sentences, as this would take them out of their context.

Eye-tracking literature has shown that gaze data could be used to account for many phenomena related to increased linguistic complexity, e.g. ambiguous words and phrases, complex syntax, unfamiliar words, etc. Based on this evidence, an ideal dataset for evaluating readability would consist of both text passages with known complexity for the target population and of gaze data obtained from these readers. The latter would allow for investigation of particular areas of difficulty within the texts and could be used to determine the level of difficulty of particular sentences as they appear in context.

Furthermore, gaze data could be used to investigate how attention works during the process of reading and thus could give useful additional insights into the evaluation of text presentation and the way users look for information within texts or within web pages.

Based on the extensive literature review presented in this chapter, we can conclude that the following are the best approaches to the assessment of text and web accessibility for people with autism:

- A readability model for people with autism should account for their difficulties in resolving ambiguity in meaning; processing text lexically and syntactically; identifying pronoun referents; making pragmatic inferences; and, ideally, comprehending figurative language.

- Readability would best be modelled using supervised machine-learning techniques, as these have been shown to outperform readability formulae.

- A combination of shallow features with more advanced syntactic, discourse and cohesion features has the potential to give the most accurate assessment of readability.

- The readability model should be evaluated on a dataset of texts with known difficulty levels for readers with autism. Such a dataset does not currently exist; hence, one should be developed.

- The complexity of individual sentences as they appear naturally in

context could be evaluated using gaze data, as gaze fixations have been shown to account for many areas of linguistic complexity.

- Gaze data could also be used for the evaluation of text presentation and for identifying the scan paths of users while looking for information within the text or the screen (e.g. in processing web pages).

The next chapter will present the development of our criterion passages, which are used for evaluating the readability classifier.

Chapter 3

Development of the ASD Corpus

3.1 Chapter Overview

This chapter presents a collection of texts whose readability was evaluated by participants with and without autism. This collection will henceforth be referred to as the ASD corpus. The development of the ASD corpus is part of RQ1 and is considered to be the first original contribution of this thesis:

RQ1: How can we obtain a collection of texts with known levels of difficulty for readers with autism?

The ASD corpus consists of 27 individual documents (4,658 words in total) of which the readability was evaluated by 27 different people who had all been formally diagnosed with autism (texts 1-16 by 20 people, texts 17-24 by 18 people and texts 25-27 by 18 people). It also contains gaze data collected while the participants were reading the texts. Parts of this chapter have been presented in Yaneva & Evans (2015) and in Yaneva et al. (2016).

3.2 Purpose of the ASD Corpus

The development of the ASD corpus has a primary and a secondary purpose.

The **primary purpose** for the development of the ASD corpus is that it can be used as a set of user-evaluated unseen data for the evaluation of the document-level readability classifier, described in Chapter 4.

The **secondary purpose** of the ASD corpus is to enable the investigation of particular sentences within the texts which may pose difficulties for readers with autism. The ASD corpus is, to the best of our knowledge, the first dataset to contain gaze data obtained from people with autism while they were reading. The corresponding gaze fixations data collected from the control participants allow for comparisons of reading in participants with and in participants without autism. The gaze data were used to determine the level of difficulty of individual sentences, as described in Chapter 5.

Finally, the gaze data from the ASD corpus could be used to investigate text constructions which pose particular reading difficulties for individuals with autism. The corresponding fixations from the control group allow for comparisons between the two groups. However, since in the context of this research the ASD corpus was used for evaluation of our document-level classifier, it was not suitable to derive features based on the particular areas of difficulty the corpus contains. Doing so would have resulted in a model which would have performed very well on the ASD corpus but which would not have generalised well over other unseen data.

3.3 Method

This section describes the method used for data collection. Prior to the start of the data-collection process, ethical approval was sought; it was granted by the relevant ethics committee.

3.3.1 Design

The study involved the evaluation of the difficulty of text passages by an experimental group of adults diagnosed with autism and a control group of neurotypical (non-autistic) adults. The participants were asked to read the texts and answer three multiple-choice questions (MCQs) per text passage. While the participants were reading the texts and answering the questions, their eye movements were recorded by an eye tracker.

Once the data were collected, the text passages were classified either as *easy*, *medium* or *difficult* based on the answers to the MCQs. The texts from these three classes were later used as user-evaluated unseen data for the evaluation of the document-level readability classifier described in Chapter 4.

The gaze data collected were used to determine the level of difficulty of individual sentences within the texts based on the number of fixations per sentence, which indicate sentence complexity (Rayner & Duffy 1986) (Chapter 2). The sentences were then ranked and classified as *easy* or *difficult* and used to develop a sentence-level readability classifier (Chapter 5).

The characteristics of the texts, MCQs, participants and the apparatus are described in the subsections below.

3.3.2 Text Passages

Selection criteria

The texts included in the experiments were not complete articles but selected passages. This was done in order to lessen the amount of time and effort required from the participants (especially those with ASD) to assess all 27 texts. The selected passages are self-contained and coherent, meaning that they do not refer to information given in the rest of the article and can be comprehended independently of it. The rest of the selection criteria are outlined below:

Prior or specialised knowledge: Texts requiring a high level of prior general or specialised knowledge were discarded. Control of this variable was necessary so as to ensure that lack of comprehension would not be due to external factors such as insufficient general knowledge. Although it is hard to measure how much prior knowledge is needed for the understanding of a concept (knowing the meaning of a word, term or a named entity, can also be regarded as prior knowledge), it was ensured that, as far as possible, all facts and events in the selected texts would be non-specialised or would be explained in the text. All events are self-contained.

Controversy of the topic: Texts containing events or opinions related to religion, sexuality, violence or to other sensitive topics were not included

in the experiments.

Terminology: None of the selected texts contained highly specialised terms, unless those terms were explained in the text.

Culture: The selected materials referred to world news and events which did not require a particular cultural background in order to be successfully comprehended.

Sources: The sources of the texts were as miscellaneous as possible in order to avoid bias based on the source. They adhered to three main registers: educational, news and general-informational articles. In total eight texts were obtained from leaflets targeted at people with cognitive disabilities, seven of which were easy-to-read leaflets produced by the National Healthcare System (UK) and one was a school leaflet. School materials comprised of eight texts from the BBC-Bitesize website[1], which contains short educational articles levelled for children from the age of seven to the age of sixteen. Three texts were obtained from the VU Amsterdam Metaphor Corpus (Steen et al. 2010), three from online personal blogs, four from various UK newspapers and one from the novel "Sense and Sensibility" by Jane Austen.

Characteristics of the selected texts

A total of 27 text passages with varying complexity were obtained from the web. The genres were miscellaneous, covering educational (seven documents), news (ten documents) and general articles (three documents), as

[1]BBC-Bitesize. available at: http://www.bbc.co.uk/education [online] [Last accessed: 08/06/2016]

Table 3.1: Characteristics of the ASD corpus

Text	**Genre**	**Words**	**FKGL**	**Flesch**
T1	Educational	163	4.93	79.548
T2	Educational	178	4.671	80.22
T3	Educational	206	7.577	65.437
T4	Educational	189	9.276	56.758
T5	Newspaper	226	11.983	40.658
T6	Newspaper	160	8.866	59.82
T7	Newspaper	163	8.765	66.657
T8	Newspaper	185	14.678	45.34
T9	Newspaper	188	9.823	58.298
T10	General	108	4.243	82.305
T11	General	141	4.561	79.108
T12	Newspaper	166	10.344	57.859
T13	Educational	209	6.087	70.124
T14	Educational	151	5.783	60.258
T15	Educational	158	6.102	57.2013
T16	Newspaper	198	13.204	46.481
T17	General	147	11.035	51.965
T18	Newspaper	227	10.171	49.093
T19	Newspaper	242	7.812	67.79
T20	Newspaper	150	9.523	64.953
T21	Easy-read	77	8.16	60.11
T22	Easy-read	96	6.73	67.33
T23	Easy-read	74	2.71	92.54
T24	Easy-read	178	5.52	75.33
T25	Easy-read	77	5.79	70.67
T26	Easy-read	121	1.75	95.00
T27	Easy-read	58	6.63	68.16

well as easy-to-read texts (seven documents). The mean number of words per text was m = 156 with standard deviation SD = 49.94. The mean number of sentences per text was m = 10.15, SD = 3.6. The texts covered a range of readability levels, where the average was m = 65.07 with SD = 13.71 according to the Flesch Reading Ease (FRE) score (Flesch 1949), which is expressed on a scale from 0 to 100 (the higher the score, the easier the text). Details about the individual texts are presented in Table 3.1. The Flesch-Kincaid Grade Level (FKGL) in Table 3.1 is proportional to text difficulty. Conversely, the Flesch Reading Ease (FRE) score, which is expressed on a scale from 0 to 100, is inversely proportional to text difficulty.

Below is an example of an educational text from the ASD corpus.

> "*Before the industrial revolution in Britain, most peppered moths were of the pale variety. This meant that they were camouflaged against the pale birch trees that they rest on. Moths with a mutant black colouring were easily spotted and eaten by birds. This gave the white variety an advantage, and they were more likely to survive to reproduce. Airborne pollution in industrial areas blackened the birch tree bark with soot. This meant that the mutant black moths were now camouflaged, while the white variety became more vulnerable to predators. This gave the black variety an advantage, and they were more likely to survive and reproduce. Over time, the black peppered moths became far more numerous in urban areas than the pale variety.*"

Another example of a text from the ASD corpus this time from a news article, follows:

> "*The season finale of The Great British Bake Off was the third most popular programme on television last year outflanked only by two World Cup football matches. The final episode of this season, airing tomorrow, will in all likelihood be the most-watched*

> *show of 2015. Over the last five years, in fact, Bake Off has so thoroughly entangled itself with the consciousness of the nation that it has become easy to forget how very, very strange it is that 10 million Britons switch on their TV sets each Wednesday evening to watch a baking contest filmed in a tent in the countryside. No one predicted the scale of its success. Richard McKerrow and Anna Beattie, who founded Love Productions, which makes the show, tried to sell the idea for four years before BBC2 finally picked it up. Their original inspiration, they told me, was the rural baking competition at a village fete; they liked the idea that bakers were naturally generous making delicious things for others.*"

Finally, an example of a general-informational text from the ASD corpus is presented below:

> "*Secondhand smoke (SHS) comes from burning cigarettes, pipes, or cigars. That smoke has many chemicals in it. Experts say that breathing SHS can harm a person's body. It can also cause headaches and make some illnesses worse. People breathe secondhand smoke when that smoke is close by. Use this countdown to help you breathe cleaner air! 1. Open a window to get some fresh air. 2. Tell the smoker how smoking affects them and YOU! 3. SHS bothers the eyes by making them burn and feel dry. 4. SHS raises the chances of getting lung diseases.*"

3.3.3 Choice of Evaluation Technique

Several techniques for measuring the level of reading comprehension have been used in readability research. The most popular of them involve using Multiple Choice Questions (MCQs), Cloze procedure (Taylor, 1953), measuring reading time and learning time (DuBay 2008). When a technique to test the readability of texts according to readers on the autistic spectrum was chosen, several important considerations were taken into account.

First, measurement of learning time was discarded as an evaluation technique in the context of this research, as a task based on measuring learning time would have placed too great a burden on the participants. Reading time was recorded and analysed; however, it was reported as an additional measure only, because it does not reflect whether the text was comprehended or not.

The Cloze test (Taylor 1953) is a type of an evaluation technique, which requires the reader to fill in missing words in the text. Using the Cloze procedure appeared to be as good a way as possible to measure the readability of the texts in this study, owing to its simplicity, to the fact that it allows researchers to measure the exact phenomenon which causes difficulty and to the argument that "unlike multiple-choice tests, cloze tests can provide suggestive information about individual sentences, clauses, phrases, and words" (DuBay 2008). However, when designing a reading-comprehension test for people with ASD, one has to take into account their good understanding of syntactic context. Frith and Snowling (Frith & Snowling 1983) report that in a cloze-test task, readers with ASD pick syntactically well-matched words, but they have an impaired understanding of semantic context, as these words are syntactically appropriate but semantically inappropriate (Frith & Snowling 1983). Using the clues provided by syntax, rather than by using comprehension alone, to help fill in the gaps could lead to a large number of deceptively correct answers (in fact, guesses) and to bias in the end results.

Unlike the cloze test, MCQs do not risk giving hints through syntactic

cues (Gronlund 1982). In this form of evaluation, the participant is asked to recognise the right answer to a question over other suggested wrong answers (distractors). Also, unlike true/false or yes/no questions, the right answer in MCQs is harder to guess and the proportion of guesses versus informed answers can be balanced through extending the number of MCQs or through penalising wrong answers (in order to keep the instruction simple, wrong answers were not penalised in this study). Results from MCQs and Cloze testing procedures are shown to have a high correlation (e.g. .76, .84, .86) when tested on various groups of readers (Bormuth 1967). However, in addition to the lack of syntactic cues in MCQs, this procedure has several other advantages over cloze testing. One of them is that unlike cloze tests, where the difficulty of items is imbalanced (Davis 1946), MCQs offer a balanced approach towards the measurement of reading comprehension, as well as that of different types of comprehension (e.g. literal versus inferred meaning), which is of significant interest when examining text comprehension in people with ASD.

To conclude, in comparison to other testing procedures, MCQs satisfy the criteria of this study best. First of all, they have better validity than measuring reading time as they account for the effects of pragmatic impairment in ASD. They have a slight advantage over the cloze procedure, as they do not influence the result by offering syntactic clues. As shown in the next section, MCQs allow testing of the level of various types of comprehension. Last but not least, MCQ tests are a popular assessment tool in schools and

thus our readers are familiar with them.

A drawback of MCQs is that "because test items themselves represent a reading task for the student, it is uncertain whether it is the difficulty of the passage or the difficulty of the items that is measured by this procedure" (Bormuth 1967). As this argument is a valid one, especially as concerns readers with comprehension difficulties such as people with ASD, we aimed to include MCQs which are as simple as possible and do not contain many clauses or complex words.

The next subsection describes the design of the multiple choice questions used in this study.

3.3.4 Design of the Multiple-Choice Questions

Since people with ASD are generally known to understand many parts of what they read literally (Happé & Frith 2006, Happe 1997, Frith & Snowling 1983, O'Connor & Klein 2004, Martos et al. 2013), it is of interest to examine different types of comprehension of the texts in the ASD corpus. Impairment in specific types of reading comprehension merits the exploration of readability features related to those specific types. An example of such a relation is the relation between the ability to make inferences and various features of cohesion explored by the cognitive paradigm in readability research (Chapter2).

Various kinds of reading comprehension have been extensively studied by Pearson & Johnson (1978), Nuttall (1996), Day & Park (2005) and others. Table 3.2 shows the main types of comprehension we examine in our study,

as well as their relation to the reading profile of people with autism.

These types of reading comprehension were examined through the inclusion of three multiple-choice questions per text passage, each of which contained three possible answers. The seven easy-to-read texts included in the ASD corpus were only examined through one literal MCQ per text, owing to the simplicity of information contained in them, which, by definition, does not require the reader to reorganise the information or make gap inferences.

An example of a multiple-choice question examining literal understanding:

Before the industrial revolution in Britain most peppered moths were:

a) Black

b) Eaten by birds

c) Of the pale variety

An example of a multiple-choice question examining inferential understanding:

Black peppered moths became more numerous in urban areas because:

a) They were mutants

c) They were camouflaged due to the airborne pollution

d) Because the airborne pollution blackened the white moths with soot

After completion of the data collection the 27 texts were divided into three classes of difficulty according to the answers given to the multiple-choice questions (Section 3.4). These three classes were later used as unseen user-

Table 3.2: Types of comprehension examined and their relation to ASD

Comprehension	Characteristics	Relation to ASD
Literal	Understanding of the straightforward meaning of the text: facts, vocabulary, dates, times, etc (Day & Park 2005)	Readers with ASD have predominantly literal understanding of language (MacKay & Shaw 2004).
Reorganisation	The ability to combine *explicitly* given information from different parts of the text. Example: *"Maria Kim was born in 1945"*; *"Maria Kim died in 1990"*. How old was Maria Kim when she died?" (Day & Park 2005).	Since this type of question is based on literal understanding it could provide insights exclusively into the roles of context and text structure, which are known to pose difficulties for people with ASD (O'Connor & Klein 2004, Oliver 1998).
Inference	The ability to use two or more pieces of information to arrive at a third piece of information that is *implicit*. Example: *"He rushed off, leaving his bike unchained"*. Inference: He left his bicycle vulnerable to theft (Kispal 2008).	Types of inferences challenging for ASD: Inferring given or presupposed knowledge as well as new or implied knowledge derived from mental state words, bridging inferences, figurative language, speaker's intention (Dennis et al. 2001)

evaluated data for the external evaluation of our document-level classifier.

3.3.5 Participants

The reading-comprehension experiments involved two groups of participants. The experimental group consisted of adult readers with a confirmed diagnosis of autism and the control group consisted of adult participants without a diagnosis of autism. The inclusion and exclusion criteria for the two groups are presented below.

Experimental group (participants with ASD):

1. **Inclusion criteria**

- Formal clinical diagnosis of Autism Spectrum Disorder, Asperger Syndrome or Pragmatic Communication Disorder
- Above 18 years old
- Native speakers of English
- Ability to read
- Minimum twelve years spent in formal education

2. **Exclusion criteria**

- Formally diagnosed developmental delay (intellectual disability)
- Comorbid disorders affecting reading (e.g. dyslexia, Attention Deficit Hyperactivity Disorder, memory disorders, etc.)
- Impaired vision (necessary for the eye-tracking experiments). All participants are required to have normal or corrected vision.

Control group (neurotypical participants):

1. **Inclusion criteria**

- Above 18 years old
- Native speakers of English
- Ability to read
- Minimum twelve years spent in formal education

2. **Exclusion criteria**

- Formal clinical diagnosis of Autism Spectrum Disorder, Asperger Syndrome or Pragmatic Communication Disorder
- Formally diagnosed developmental delay (intellectual disability)
- Comorbid disorders affecting reading (e.g. dyslexia, Attention Deficit Hyperactivity Disorder, memory disorders, etc.)
- Impaired vision (necessary for the eye-tracking experiments). All participants are required to have normal or corrected vision.

Data were collected from both groups of participants; however, the texts in the ASD corpus were divided into three groups of *easy, medium* and *difficult* based on the answers of the participants with ASD.

The evaluation of the texts was performed in three cycles of data collection conducted in the span of two years and involved 27 different participants with a confirmed diagnosis of autism.

Texts 1-9 and 21-27 were evaluated by Group 1, consisting of twenty adult ASD participants (thirteen male, seven female) with mean age (m) in

years m = 30.75 and standard deviation SD = 8.23; years spent in education, as a factor influencing reading skills, were m = 15.31, with SD = 2.9. The control group participants who evaluated these texts were twenty non-autistic adults (eleven female and nine male), with mean age m = 30.81, SD = 4.8 and years spent in education m=17.25, SD=2.15.

Texts 10-17 were evaluated by Group 2, consisting of eighteen adult participants with ASD (eleven male and seven female) with mean age m = 36.83 , SD = 10.8 and years spent in education m = 16, SD = 3.33. The control-group participants were eighteen adults (twelve male and six female) with mean age m = 33.11, SD = 8.19 and years spent in education m = 17.89, SD = 3.55.

Texts 18-20 were evaluated by Group 3, which consisted of eighteen adults with autism (twelve male and six female) with mean age m = 37.22, SD = 10.3 and years spent in education m = 16, SD = 3.33. The control group participants were fourteen adults (nine male and five female) with mean age m = 34.5 , SD = 8.19 and years spent in education m = 18.93, SD = 3.1.

All participants were native speakers of English. None of them had other conditions affecting reading (e.g. dyslexia, intellectual disability, aphasia etc.). Some participants were diagnosed also with depression (n=4, ASD group; n=1, control group) and anxiety (n=6, ASD group).

3.3.6 Apparatus

The device used for recording the gaze of the participants was a Gazepoint GP3 video-based eye tracker (Gazepoint 2015) (60Hz sampling rate and accuracy of 0.5 - 1 degree of a visual angle). Texts were presented on a 19" LCD monitor. The eye tracker was calibrated individually for each participant using a 9-point calibration procedure. The distance between each participant and the eye tracker was controlled using a sensor integrated within the Gazepoint software, and was approximately 65 cm. The software randomised both the order of presentation of the texts and the questions pertaining to texts for each participant, in order to avoid bias.

3.3.7 Procedure

Instruction. Participants were informed what the purpose of the experiment was and that their individual answers would not be made public. Each participant was given the opportunity to ask questions and to request a break at any point if he or she felt tired. Recalibration was performed if the participants needed to get up during their breaks. Participants were instructed that they were free to withdraw from the research at any time. The participants were also informed of the use of an eye-tracking device: what it was and by what process eye fixations were recorded. Each participant signed a consent form and retained an information sheet, explaining all relevant information.

Collection of demographic data. Each participant was asked for his or her age, gender, formally diagnosed conditions and years of schooling.

Calibration of the eye tracker. A nine-point calibration was performed for each participant. He or she could have as many attempts at calibrating the device as necessary.

Comprehension testing. The presentation of texts was on a 19" LCD monitor, using a simple ASD-friendly interface, specifically designed[2] for the purpose of the task (not containing logos or other distracting information, having only one navigation button and a simple navigation procedure). First, a random text was presented on the screen. When participants had finished reading it, they pressed the "Enter" button, which took them to the first question for the text (both the order of the texts and the questions were randomised). Each question and its possible answers were presented at the top of the screen, with the text the question referred to displayed beneath. This was done in order to eliminate memory as a confounding variable. After the participant had completed the first question, he or she pressed "Enter" and proceeded to the second question followed by the third one. After that they proceeded to the next text and so on, until all texts were evaluated.

All participants completed the experiment in a quiet room with only the researcher present.

Control of variables. In order to maximise the internal validity of the

[2]We thank Dr. Miguel Angel Rios Gaona for his valuable help with the design of the software, which displayed and ransomised the texts and the questions.

experiment, the texts were presented in random order to each participant. This controlled for factors such as fatigue or participants becoming accustomed to the types of questions. The order of questions after each text was also randomised, so that it would not influence the answers given by the participants. The effects of memory were controlled by having the relevant passage constantly displayed on the screen. Participants could therefore refer to it whenever they were not sure about the information it contained. While the effects of background knowledge could not be eliminated entirely, the selection of texts was made in such a way, as to ensure that this effect would be minimised as far as possible. All tests were performed in a quiet room and a relaxed atmosphere, minimising the influence of environmental stressors. Control of head movements for the eye tracking experiment was provided through detailed instruction. No objects were used to restrict the head (e.g. a chin rest), due to the anxiety and sensory issues they could cause in people with autism.

Debriefing. At the end of the experiment participants were debriefed and a debriefing sheet was provided for them to keep.

This section has presented the process of data collection for the ASD corpus. The sections below describe the process of classifying the text passages into three levels of difficulty and the processing of the eye-tracking data for analysis of individual sentences.

3.4 Classification of the Text Passages into *Easy, Medium* and *Difficult*

The 27 texts from the ASD corpus were divided into classes of *easy*, *medium* and *difficult* based on the answers to the multiple-choice questions given by the participants with autism. First, each text was evaluated using three MCQs and each correct answer was given one point, while each incorrect answer was given zero points. Thus, if a participant had answered two out of three questions correctly for a given text, then that text had an answering score of two for this participant. After that all answering scores for the participants were added for each text, the texts were ranked and split into three groups impressionistically based on this ranking. The differences in the levels of difficulty of these three groups was tested as follows.

A Shapiro-Wilk test showed that the data was non-normally distributed; hence, a Friedman test was applied, which established that there were significant differences between the three groups of answer scores obtained by the ASD group ($\chi^2(16) = 134.690$, $p = 0.000$). In order to compare the three groups of texts individually, we applied a Wilcoxon Signed rank post-hoc test with Bonferroni corrections to the significance level ($\alpha = 0.017$). The results from the Wilcoxon test indicated that the texts classed as *difficult* were significantly more complex than the texts classed as *medium* (Z = -5.762, $p = 0.000$) and those classed as *easy* (Z = -9.479, $p = 0.000$) and that those classed as *medium* were significantly more complex than those classed

as *easy* (Z = -6.350, $p = 0.000$). Thus, ten texts were classified as *easy*, eight texts were classified as *medium* and nine texts were classified as *difficult*.

3.5 Processing of the Eye-tracking Data

This section presents the processing of the data collected via the eye-tracking recordings and the production of a version of the ASD corpus with assigned gaze-data metrics for each word.

3.5.1 Ensuring the Quality of the Gaze Data

All gaze data were manually corrected for vertical systematic error by following a procedure recommended in Hornof & Halverson (2002), where we used navigation buttons as required fixation locations. These locations served as reference points when examining the possible dispositioning of the gaze path (e.g. reading above the line). Data inaccuracies usually result from poor calibration or system imprecisions typical for each eye tracker (Duchowski 2009); however, in the case of the autistic participants inaccuracies also resulted from too many head movements and reduced ability to follow instruction (Sasson & Elison 2012). Owing to this added procedural difficulty, data which could not be corrected because the error was not systematic (e.g. due to too many head movements or an inability to calibrate the eye tracker) were not included in the study. Thus, the final number of participants from whom the gaze data was retained were nine ASD and nine control participants for

texts 1-9, thirteen ASD participants and fourteen control participants for texts 10-17 and finally, eight ASD participants and ten control participants for texts 18-19. Gaze data from the easy-to-read documents were not included in the analysis as they were initially collected following a different task which resulted in more fixations.

Figure 3.1 shows the scan path before the correction of vertical systematic error and Figure 3.2 shows the corrected scan path, where the gaze fixations fall onto the text lines.

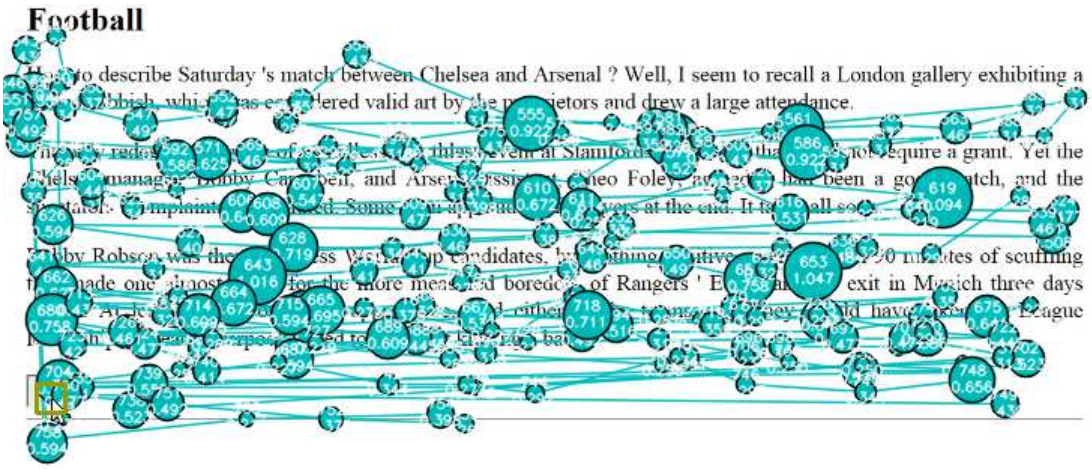

Figure 3.1: Gaze path *before* the correction of vertical inaccuracy

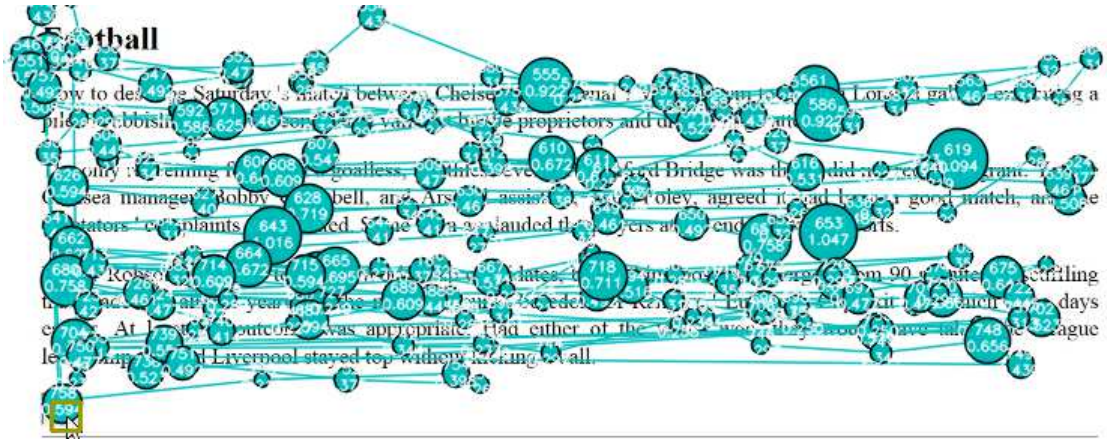

Figure 3.2: Gaze path *after* the correction of vertical inaccuracy

3.5.2 Part-of-speech Tagging and Assigning Gaze Metrics to Individual Words

Each word from the texts was defined as an Area of Interest (AOI), as shown in Figure 3.3; in total there were 3636 AOIs. Three gaze-based metrics were computed for each AOI using Gazepoint analysis software (Gazepoint 2015):

Average Time Viewed (ATV): The average time an AOI was viewed by all participants in a group (ASD or control) measured in seconds.

Average Number of Fixations (AF): The average number of gaze fixations from all participants in a group (ASD or control) in a given AOI.

Average Number of Revisits (AR): The average number of times participants went back to a previously viewed AOI. This measure is particularly relevant to measuring the heavy cognitive load posed by particular text constructions and is informative about the process of information integration

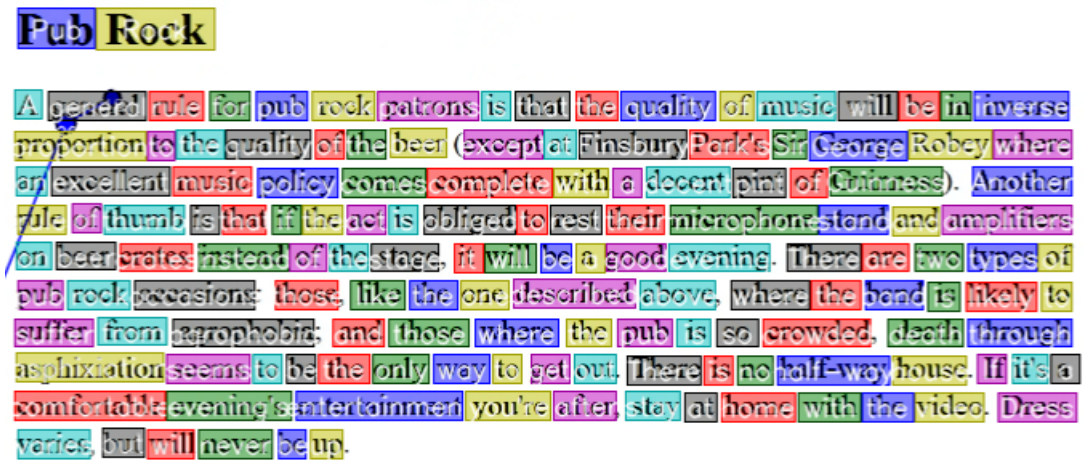

Figure 3.3: Areas of interest for each word in the text

in the readers.

All texts were processed using the Stanford parser[3] (Klein & Manning 2003). The resulting corpus is a .csv file containing all eye-tracking data from both groups, Part-of-Speech (POS) tags for each word, and anaphoric links within the texts, as shown in Table 3.3.

The resulting corpus of paired gaze data and comprehension scores could be used for the investigation of the differences between the two groups of readers, for research from the perspective of clinical linguistics or to investigate which linguistic phenomena impose greater cognitive load on participants with autism and those without autism. For example, figure 3.4 illustrates the effect of word complexity on gaze fixation duration, where the word "sonorous" (positioned last) has been fixated longer than any other word in

[3]Available at: http://nlp.stanford.edu/software/lex-parser.shtml

Table 3.3: An example of the corpus data obtained from participants with autism (A) and neurotypical control participants (C)

Item	AOI	POS	Coref	A-ATV	A-AF	A-AR	C-ATV	C-AF	C-AR
14	Your	prp$	set 11	0.225	2.229	2.618	0.221	2.234	2.505
15	team	nn		0.22	2.219	2.447	0.213	2.075	2.076
16	is	vbz		0.112	1.704	1.959	0.108	1.859	2.024
17	losing	vbg		0.255	2.155	2.438	0.297	2.4	72.89
18	by	in							
19	just	rb		0.198	1.833	2.094	0.194	1.788	2.067
20	one	cd		0.159	1.945	1.945	0.149	1.762	2.051
21	goal	nn		0.188	1.903	1.852	0.184	1.966	2.789
22	.	.							

the sentence *"they do not make a ringing sound when they are hit (they are not sonorous)"*.

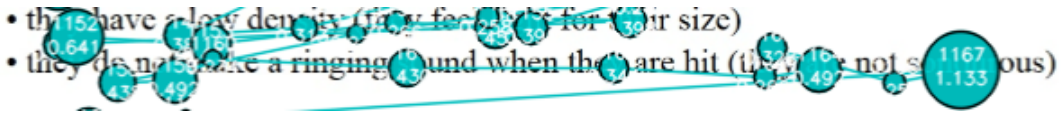

Figure 3.4: Example of the effect of word complexity on gaze fixation duration (in this case, the complex word "sonorous")

This section has described the processing of the gaze data for the ASD corpus. The next section discusses the methodological challenges encountered during the development of the corpus, and the limitations and contributions of the corpus.

3.6 Discussion

This section discusses the contributions and limitations of the ASD corpus and the challenges encountered during the development of it.

3.6.1 Methodological Challenges and Contributions

The development of the ASD corpus described in this chapter is considered to be the first original contribution of this thesis:

Contribution 1. The development of a corpus of texts with known reading difficulty for readers with autism (the ASD cor-

pus)(RQ1)

By far the biggest methodological challenge for the development of the corpus was the recruitment of a sufficient number of participants with autism, who fulfilled the inclusion criteria but did not fall under the exclusion criteria presented earlier in this chapter. This difficulty was even greater considering the challenges some people with autism have with comprehending instructions, which in some cases forced us to discard some of the data because of inaccuracies (e.g. too many head movements). However, the ASD corpus is the first dataset of its kind to contain texts with known reading difficulty for people with autism and to contain paired gaze fixations from two groups of participants (ASD and neurotypical). The development of this corpus allows for the first evidence-based evaluation of a readability classifier for people with autism (Chapter 4).

However, the data collection process and the resulting ASD corpus itself are not without limitations and these limitations need to be taken into account when designing experiments involving this data.

3.6.2 Limitations

The first limitation is the low sampling rate of the eye tracker (60Hz), as a result of which not every word in the corpus if fixated upon by all participants. In spite of the fact that extra care was taken to remove vertical systematic error, it is still possible that some of the fixation locations may not be as accurate as in other eye-tracking corpora of data obtained using

faster devices. Thus, while the data are still useful for the analysis of larger text entities such as sentences, the corpus should be applied with caution to linguistic phenomena which are more fine-grained (e.g. short words) or which appear with low frequency within the corpus, as there may not be a sufficient number of fixation points (e.g. for figurative expressions).

Another limitation is the small number of participants involved in the assessment of the text. One reason for this is the fact that the participants had to be recruited from among those with a diagnosis of autism instead of simply from a general population of readers. Furthermore, we applied robust exclusion criteria, which excluded all people on the autism spectrum who had any form of intellectual disability. This was necessary in order to ensure that any comprehension difficulties in the experiments were caused by the presence of autism and not of intellectual disability, but on the other hand, this limited the recruitment to a small subset of people on the autism spectrum who had high-functioning autism.

Another limitation is the small size and the low number of text passages in the corpus. This was necessary in order to avoid fatigue in the participants and to comply with ethical considerations. For comparison, LocalNews (Feng 2009), which is the only other readability corpus for English evaluated by people with cognitive disabilities, features 11 original and 11 simplified texts.

Last but not least, a large portion of the recorded gaze data was discarded (from eleven participants from Group 1 (55%), from five participants from Group 2 (27.7%) and from ten participants from Group 3 (55%)). This

was due to system inaccuracies, unsuccessful calibration of the eye tracker in some participants and too many head movements. However, discarding these data was necessary in order to make sure that the analysed data was of good quality and contained as few noise fixations as possible. Finally, some inaccuracies in the annotation of anaphora or part-of-speech may have resulted from the use of an automatic parser.

All limitations listed above should be accounted for when experiments using this corpus are designed.

3.7 Summary

This chapter discussed the development of the ASD corpus. First, we presented the design of the experiments for evaluating the difficulty of text passages by two groups: participants with and participants without autism. We then discussed the selection and characteristics of the text passages and of the multiple-choice questions used in the experiments, the inclusion and exclusion criteria for the participants, as well as the apparatus and procedure employed in the study. We then described how the text passages were classified into groups of *easy, medium* and *difficult* texts according to the answers of the participants with autism. This was followed by a description of the gaze-data processing. Finally, the advantages and limitations of the resulting ASD corpus were discussed.

The next Chapter 4 will present the development and evaluation of a

document-level readability classifier for people with autism.

Chapter 4

Document-level Readability Assessment

4.1 Chapter Overview

This chapter describes the corpora, features and algorithms used in the development of an automatic document-level readability classifier for readers with autism. The development of this classifier addresses research question number two:

RQ2: Is it possible to develop an automatic *document*-level readability classifier for people with autism, that generalises over unseen user-evaluated data better than existing readability metrics?

The generalisability of the classifier is evaluated on the ASD corpus (Chapter 3) and is compared to a common baseline.

4.2 Purpose of the Document-level Classifier

The purpose of this classifier is to help professionals who develop texts that are accessible to readers with autism to evaluate the readability of their output without needing access to a focus group of people with cognitive

disabilities. It can also be used to evaluate the output of automatic text simplification systems.

4.3 Corpora

This section describes the corpora used for the training and intrinsic evaluation of the classifier, as well as the evaluation of its generalisability.

4.3.1 Training Corpus

The corpus used for the training of the classifier was the WeeBit readability corpus (Vajjala & Meurers 2012). The WeeBit corpus contains articles obtained from the *Weekly Reader* [1] and educational articles from the BBC-BiteSize[2] website. The WeeBit corpus comprises two sub-corpora of the same names.

The articles from the *Weekly Reader* cover a wide range of topics, from non-fiction to current affairs, and were downloaded in November, 2011 (Vajjala Balakrishna 2015). The images and weekly quizzes contained in the online magazine were not featured in the corpus, so it only contains articles. The *WeeklyReader* is aimed at children of ages 7-8 (Level 2), 8-9 (Level 3), 9-10 (Level 4) and 9-12 (Senior level). The criterion for the evaluation of the graded writing has not been published by the magazine (Vajjala Balakrishna

[1] http://www.weeklyreader.com/

[2] http://www.bbc.co.uk/education

2015).

BBC-BiteSize is also an educational site containing articles at four levels corresponding to educational key stages (KS) for children between ages 5-7 (KS1), 7-11 (KS2), 11-14 (KS3) and 14-16 (GCSE). The articles from this website which are featured in the WeeBit corpus were downloaded in 2009.

The combined WeeBit corpus comprises five readability levels corresponding to the *Weekly Reader's* Level 2 (807 documents, Class 1), Level 3 (789 documents, Class 2) and Level 4 (629 documents, Class 3) and BBC-BiteSize KS4 (646 documents, Class 4) and GCSE levels (7530 documents, Class 5). The average document length is 23.4 sentences at the lowest level and 27.8 sentences at the highest level (Vajjala Balakrishna 2015).

As the purpose of our work is to build a three-level readability classifier for people with autism, we first balanced the number of documents per class and then normalised the WeeBit corpus to include texts of only three readability levels: *easy*, *medium* and *difficult*. Thus, from Classes 1-5 we excluded classes 2 and 4 and retained Class 1 (807 documents, *easy*), Class 3 (629 documents, *medium*) and Class 5 (balanced to 703 randomly selected documents, *difficult*), leaving 2139 documents in total. Table 4.1 presents the total classes of the Weebit corpus, where the ones marked in bold were selected to represent the training set for our classifier.

The WeeBit corpus was used for the training and intrinsic evaluation of the document-level readability classifier.

Table 4.1: The WeeBit corpus (Vajjala & Meurers 2012). The classes marked in bold were used for training of our document-level classifier

Level	WeeBit Class	Ages	Total texts
1	**Level 2**	**7-8**	**807**
2	Level 3	8-9	789
3	**Level 4**	**9-10**	**629**
4	KS3	11-14	646
5	**GCSE**	**14-16**	**7530**

4.3.2 Evaluation Corpus

After the classifier was trained and intrinsically evaluated on the Weebit corpus, its generalisability was tested on a set of unseen user-evaluated data, namely the texts from the ASD corpus described in Chapter 3.

4.4 Features

A total of 43 individual features were employed in the development of the document-level readability classifier. These features were categorised into 1) lexico-semantic, 2) syntactic, and 3) cognitively-motivated features, 4) features of cohesion and 5) readability formulae. The cohesion features and cognitively motivated features were inspired by the Coh-Metrix tool (McNamara et al. 2014).

4.4.1 Lexico-semantic Features

This group includes surface lexical features such as *Number of syllables in long words* and *Average word length in syllables*, as well as features reflecting various semantic aspects of the words such as polysemy (e.g. *Number of polysemous words*) or lexical diversity (e.g. *Type-token ratio*). Table 4.2 presents a list of the lexico-semantic features used for document classification and their descriptions.

Polysemy refers to the number of senses a word has. Polysemous words are considered more difficult to process since they offer more than one possible lexical interpretation and thus introduce ambiguity (DuBay 2008). However, frequent words tend to be more polysemous than rare or specialised words (DuBay 2008), which is why the relation between polysemy and text complexity is not straightforward. Still, polysemy has been found particularly challenging for readers with autism (Happé & Frith 2006, Happe 1997, Frith & Snowling 1983, O'Connor & Klein 2004, Martos et al. 2013), which is why our document-level classifier accounts for it through semantic features such as *Number of polysemous words* and *Polysemous type ratio*. Both of these features are computed based on the polysemy relations between words in WordNet (Miller 1995). WordNet is a lexical database, which contains groups of related lexical items called synsets. Polysemy relations in WordNet are based on synsets, where a polysemous word (e.g. *"bank"*) would be assigned to more than one synset (e.g. one referring to its meaning of a

Table 4.2: Document classification: Lexico-semantic features

Feature	**Description**
Long words	Proportion of words in the text with 3 or more syllables
Average word length	Average number of syllables for all words
Number of polysemous words	Words with more than one sense in WordNet
Polysemous type ratio	Ratio of polysemous types to word types (content words)
Type-token ratio	Total number of types/number of tokens (content words)
Vocabulary variation	Word types divided by common words not in the text
Numerical expressions	Number of numerical expressions
Number of infrequent words	Words not among the 5,000 most frequent words in English
Total number of words	Total number of words in the text
Dolch-Fry Index	Words in the *Fry 1000 Instant Word List/Dolch Word List*
Number of passive verbs	Number of passive verbs
Agentless passive density	Incidence score of passive voice
Negations	Number of negations
Negation density	Incidence score of negations

financial institution and another referring to its meaning of a river side).

Another characteristic of a text, which determines its complexity is lexical diversity. Lexical diversity refers to the relationship between the number of unique words in the text (types) and the total number of words in the text (tokens). A high number of different words in a text indicates that new words need to be integrated into the discourse context, which increases its difficulty (McNamara et al. 2014). In this thesis, lexical diversity is measured by *Type-token ratio*, *Vocabulary variation* and *Number of numerical expressions* (Table 4.2).

Some statistical measurements such as *Number of infrequent words* and *Total number of words* were also included as features in the classifier, since short texts which contain common words are easier to comprehend than long texts containing rare words. Word frequency was also measured using the *Dolch-Fry Index*, which evaluates the proportion of words in the text that appear in the *Fry 1000 Instant Word List* (Fry 2004) or the *Dolch Word List* (Dolch 1948).

The cognitive load imposed by lexico-semantic processing was also measured through features such as *Number of passive verbs, Agentless passive density, Negations* and *Negation density* (Table 4.2).

4.4.2 Syntactic Features

The syntactic complexity of a text is associated with delayed processing time and understanding (Gibson 1998) and is known as a source of significant challenges to readers with autism (Whyte et al. 2014).

To account for the syntactic complexity of texts, we included surface features such as *Long sentences*, *Words per sentence*, *Average sentence length, Total number of sentences* and *Paragraph index*. In addition, features quantifying the number of punctuation marks indicating syntactic complexity were evaluated: *Number of semicolons/suspension points, Number of Unusual punctuation marks* and *Comma index*. Table 4.3 presents the syntactic features used for document classification.

Table 4.3: Document classification: Syntactic features

Feature	**Description**
Long sentences	Proportion of sentences longer than 15 words
Words per sentence	Total words / total sentences
Average sentence length	Sentence length in words
Total number of sentences	Total number of sentences
Paragraph index	10 x total paragraphs / total words
Number of semicolons	Number of semicolons
Number of Unusual punctuation marks	Number of occurences of &, %,
Comma index	10 x total commas / total words

4.4.3 Features of Cohesion

Increased cognitive load is not related solely to the lexical properties of the text, but also to the way in which content is organised. Cohesion has been defined as "a phenomenon accounting for the observation (and assumption) that what people try to communicate in spoken or written form under 'normal circumstances' is a coherent whole, rather than a collection of isolated or unrelated sentences, phrases or words" (Halliday & Hasan 1976). Cohesion involves different types of relationships within the text, such as temporal or causal relationships, which are normally represented by specific types of connectives like temporal conjunctions (e.g. *first, until*) or causal conjunctions (e.g. *because, so*). We evaluate referential cohesion (overlap in content words between local sentences) (McNamara et al. 2014) and overall discourse cohesion. Referential cohesion is measured by computing incidence

Table 4.4: Document classification: Features of cohesion

Feature	**Description**
Pronoun Score	Occurence of pronouns per 1,000 words
Definite description score	Occurence of def. descriptions per 1,000 words
Number of illative conjunctions	Number of illative conjunctions
Number of comparative conjunctions	Number of comparative conjunctions
Number of adversative conjunctions	Number of adversative conjunctions

scores (occurrence per 1000 words) of *Pronouns* and *Definite descriptions*. Discourse cohesion is measured by computing incidence scores of *Numbers of illative conjunctions* (e.g. *for, so*), *Comparative conjunctions* (e.g. *just as, likewise*), *Adversative conjunctions* (e.g. *although, whereas*). Table 4.4 presents a description of each of the features of cohesion used for document classification.

The computation of the cohesion features was inspired by Coh-Metrix (McNamara et al. 2014) and the code for them was re-implemented following the definitions of the features in McNamara et al. (2014).

4.4.4 Cognitively-motivated Features

The source for the cognitively-motivated features used in this research was a set of word lists from the MRC Psycholinguistic Database (Coltheart 1981). Each word in these lists has an assigned score based on human rankings, obtained by presenting stimulus words to large numbers of subjects who are then asked to evaluate word properties such as how easy or difficult it is to

Table 4.5: Document classification: Cognitively-motivated features

Feature	**Description**
Word frequency	Average frequency of words
Age of acquisition (average)	Age of acquisition norms from the MRC database
Familiarity (average)	Familiarity norms from the MRC database
Concreteness (average)	Concreteness norms from the MRC database
Imagability (average)	Imagability norms from the MRC database
Number of 1st pronominal reference	Number of 1st pronominal reference
Number of 2nd pronominal reference	Number of 2nd pronominal reference

picture mentally the word referent (imagability), how abstract or concrete the word is, the age at which the word is first acquired (evaluated using children), etc. (Chapter 2). The cognitively-motivated features included in the classifier were *Average word frequency, Age of acquisition, Imagability, Concreteness* and *Familiarity*.

Another two features were the *Numbers of first and second person pronominal references*, which were included in the classifier because a higher number of personal words in a text (e.g. *I, you*) is recommended as a way to improve ease of comprehension (Freyhoff et al. 1998).

Table 4.5 presents a list of the cognitively-motivated features used for document classification and their descriptions.

4.4.5 Readability Formulae

Finally, readability has been traditionally measured using various readability formulae. We have used several of them as features in the classifier, as detailed below.

Automated Readability Index (ARI) (Smith et al. 1989)

$$ARI = 4.71 \times CPW + 0.5 \times SLW - 21.43 \tag{4.1}$$

In this formula, CPW stands for characters per word and SLW stands for sentence length in words.

Coleman-Liau formula (CL) (Coleman 1971). L is the average number of letters per 100 words. S is the average number of sentences per 100 words.

$$CL = 0.0588 \times L - 0.296 \times S - 15.8 \tag{4.2}$$

Fog Index (Gunning 1952), where the grade level (GL) is determined by average sentence length (ASL) and the number of hard words (HW) for each 100 words of a document.

$$GL = 0.4 \times (ASL + HW) \tag{4.3}$$

Lix formula (Anderson 1983). In the Lix formula A = Number of words, B = Number of periods (defined by period, colon or capital first letter) and C = Number of words with more than six letters.

$$LIX = \frac{A}{B} + \frac{(C \times 100)}{A} \tag{4.4}$$

SMOG Reading Ease (McLaughlin 1969), where polysyllable count (PSC) is used (words that contain more than two syllables in 30 sentences)

$$SMOG = 3 + \sqrt{PSC} \tag{4.5}$$

Flesch Reading Ease (Flesch 1948)

$$FRE = 206.835 - 1.015 \times \frac{words}{sentences} - 84.6 \times \frac{syllables}{words} \tag{4.6}$$

Flesch Kincaid Grade Level (Kincaid et al. 1975)

$$FKGL = 0.39 \times \frac{words}{sentences} + 11.8 \times \frac{syllables}{words} - 15.59 \tag{4.7}$$

FIRST Readability Index. CI is Comma Index, PI is Paragraph Index, SI is Syllable Index, SLI is Sentence Length Index, TTR is Type Token Ratio, VV is Vocabulary Variation, and DFI is Dolch-Fry Index. The FIRST readability index was developed specifically for people with autism in the EC-funded FIRST project by professionals in mental healthcare (Jordanova et al. 2013).

$$FIRST = 95.43 - (0.076 \times CI) + (0.201 \times PI) - (0.067 \times SI) - (0.073 \times SLI)$$

$$- (35.202 \times TTR) - (1.060 \times VV) + (0.778 \times DFI) \tag{4.8}$$

4.5 Experimental Setup

This section presents the experimental setup for the training and evaluation of the document-level readability classifier: the algorithms, baseline, feature-selection process and the evaluation techniques used.

4.5.1 Modelling Method

The chosen modelling method for this task is classification. The motivation behind this choice lays in the practical purpose of the classifier. Ultimately, the model is designed to be used by humans and text simplification systems in order to provide feedback on the specific level of difficulty of their texts for readers with autism as opposed to the relative difficulty of texts compared to one another, feedback on which could be provided by a regression model.

Hence, in the results section we report the measurement of classification accuracy - i.e., how accurate the classifier is in assigning the correct class to an unseen document given as a percentage. Other reported measures are precision (proportion of returned results that are relevant), recall (proportion of relevant results that are returned), and F-measure (a harmonic mean of precision and recall).

4.5.2 Algorithms

The document-level readability classifier was built using supervised learning algorithms implemented in the Weka toolkit (Waikato Environment for Knowledge Analysis) (Frank & Witten 1998). A number of algorithms were evaluated on the WeeBit dataset. The results section below presents only those algorithms that achieved the best results in terms of accuracy when evaluated on the WeeBit corpus and on the unseen data (the ASD corpus).

4.5.3 Baseline

We use the Flesch-Kincaid Grade Level readability formula (Kincaid et al. 1975) as a baseline for document classification because it is one of the best-performing predictors of text difficulty; it is also used as a baseline in other readability estimation models (Vajjala Balakrishna 2015). The baseline values are computed by using the formula as a single feature in the classification model.

4.5.4 Feature Selection

Initially the full-feature set was used to obtain a baseline model, which was subsequently optimised through the Best First attribute selection filter for supervised learning which is built into Weka (Frank & Witten 1998).

The selected features after the Best First attribute filter was applied were: *Polysemous type ratio, Words per sentence, Fog index, Average sentence length, Age of acquisition of words, Second pronoun incidence, Imagability* and *Flesch-Kincaid Grade Level*.

4.5.5 Training and Internal Validity Evaluation

The classifier was trained and intrinsically tested using the 10-fold cross-validation evaluation strategy, in which the dataset is randomly divided into ten folds of equal dimensions. Then, each classifier is trained on nine of these folders and tested on the tenth and the process is iterated so that each fold

becomes a test fold once. Classification effectiveness is then averaged out across all test folds and reported as an F-measure.

Since this way of evaluating the classifier uses training and evaluation data from the same sample of texts (the WeeBit corpus), it is a measure of the internal validity of the classifier.

4.5.6 Generalisability

After the internal validity of the classifiers was evaluated using 10-fold cross-validation, their generalisability (external validity) was evaluated on the ASD corpus. In this context, the term generalisability refers to the success with which a classifier trained on one sample (in our case the WeeBit corpus) learns rules that perform well when applied to a broader population (in this case the unseen data from the ASD corpus).

The results for the internal and external validity of the classifier are presented in the next section.

4.6 Results

This section presents results for the document-level classifier after evaluating the performance of a number of supervised learning algorithms in Weka (Frank & Witten 1998). We present results for the two best-performing algorithms, the Random Forest algorithm and the REPTree algorithm. Random Forest achieved the best internal validity of all tested algorithms, while

REPTree generalised best over unseen data. However, the accuracy of the classifier based on the REPTree algorithm could not be determined without tuning on the test corpus.

The Random Forest algorithm (Breiman 2001) is a decision-tree algorithm which uses multiple random trees to "vote" for an overall classification of the given input. It uses the *bagging* technique, which is built on the rationale that a combination of learning models increases the classification accuracy. The name "forest" comes from the fact that the algorithm functions as a collection of decorrelated decision trees, where each tree is created based on a subset of random samples from the dataset. After that all of the decision trees are used to create a ranking of classifiers, in which each tree makes an independent decision regarding a new element (e.g. a new text). Then the votes (assigned classes) of all trees are compared and the new element is assigned the class that the majority of trees have voted for (Hastie et al. 2001).

From Table 4.6 we see that the accuracy for 10-fold cross-validation using the Random Forest classifier is 0.9 (90%) and the accuracy for the test set (the ASD corpus) is 0.78 (78%). For comparison, the baseline for 10-fold cross-validation was 0.58 (58%) and 0.52 (52%) respectively.

The Reduced Error Pruning Tree algorithm (REPTree) is also a decision-tree algorithm. It first builds a tree by calculating the information gain using entropy (Quinlan 1987, Witten & Frank 2005). Entropy is a measure of impurity and higher entropy means that there is more information

Table 4.6: Document-level classifier results for Random Forest algorithm

	10-fold Cross-Validation			**Test on the ASD corpus**		
	Baseline	All features	Sel. features	Baseline	All features	Sel. features
Precision	0.576	0.911	0.903	0.548	0.739	0.798
Recall	0.576	0.91	0.903	0.519	0.667	0.778
F	0.576	0.910	0.903	0.526	0.680	0.782
Accuracy	0.58	0.91	0.90	0.52	0.67	0.78

content. For example, a training set which contains examples from only one class will have an entropy of zero, while a training set which contains half of its examples from one class and half of its examples from another will have an entropy of one.

Once the decision tree is built the algorithm reduces the error arising from variance by using reduced error pruning (Quinlan 1987, Witten & Frank 2005). Reduced error pruning is a technique in machine learning where the size of the decision tree is reduced by replacing each node in the leaves with its most popular class. Then, prediction accuracy is tested and if it has not dropped as a result of this replacement, the change is kept.

From Table 4.7 we see that the accuracy for 10-fold cross-validation using the REPTree classifier is 0.86 (86%) and the accuracy for the test set (the ASD corpus) is 0.85 (85%). For comparison, the baseline for 10-fold cross-validation was 0.6 (60%); for the test set it was 0.67 (67%).

Table 4.7: Document-level classifier results for the REPTree algorithm

	10-fold Cross-Validation			**Test on the ASD corpus**		
	Baseline	All features	Sel. features	Baseline	All features	Sel. features
Precision	0.605	0.861	0.866	0.657	0.709	0.88
Recall	0.602	0.86	0.864	0.667	0.667	0.852
F	0.603	0.861	0.864	0.64	0.679	0.85
Accuracy	0.6	0.86	0.86	0.67	0.67	0.85

4.7 Discussion

This section discusses the main findings from the experiments on document-level classification, as well as the challenges, contributions, limitations and avenues for future research, which arise from this work.

4.7.1 Methodological Challenges and Contributions

The document-level classifiers presented in this chapter preform significantly better than the baselines for both the WeeBit and the ASD corpora. It was interesting to note that for both the Random Forest and the REPTree algorithms, there was almost no difference introduced by the selection of features for 10-fold cross-validation. However, when evaluated on unseen data, the selected-features model performed with 11% better accuracy for the Random Forest algorithm and 18% better accuracy for the REPTree algorithm. In addition, the choice of a classification algorithm also played a significant role in achieving optimal accuracy. Both of these observations

suggest that more research in domain adaptation is needed for the specific registers of texts used. In order to explore the question of data from different domains further, it would be interesting to compare models trained on large generic corpora versus models trained on small user-specific corpora.

In its current version, the document level classifier has the potential to be used as a pre-evaluation technique in the development of accessible documents or, in cases where user evaluation is not feasible, it could be the only way to evaluate the ASD-accessibility of a document. We do not recommend ad-hoc use of the tool, as the history of readability research has shown us many cases where the coherence of a text is broken so that it fits the formulae better (Chapter 2).

4.7.2 Limitations

In spite of the comparatively high accuracy achieved by the classifiers, this research is not without its limitations. The first of these limitations is related to the size of the ASD corpus (27 documents), which is too small to account fully for the great heterogeneity of natural language. A similar problem could be found in many areas of disability research, where the development of resources with the involvement of the target users is challenging, time-consuming and expensive. As already described in Chapter 3, we have attempted to include texts from miscellaneous registers in order to compensate for the small size of the corpus.

Avenues for future work include investigation of the possibility of using

gaze fixations as a proxy for measuring the complexity of an entire text by correlating the average number of fixations and answers to comprehension questions. Another way in which the usefulness of the readability estimation could be improved would be to develop text classification for people with autism which divided texts into more than three levels of complexity.

4.8 Summary

This section has presented the development and evaluation of a document-level readability classifier for readers with autism, which distinguishes between three levels of text difficulty. First, we presented the corpora used for training, intrinsic and extrinsic evaluation of the classifier. We then presented the features, modelling method and classification algorithms. Once the classifier was evaluated intrinsically using 10-fold cross-validation, its generalisability was tested on unseen data, namely the user-evaluated ASD corpus (Chapter 3).

The next section will present a more fine-grained approach to readability, where we aim to assess the readability of individual sentences.

Chapter 5

Sentence-level Readability Assessment

5.1 Chapter Overview

This chapter describes the method, data and features used in the development of an automatic sentence-level readability classifier for readers with autism. The development of this classifier addresses research question number three:

RQ3: Is it possible to develop an automatic *sentence*-level readability classifier for people with autism, that performs better than existing readability metrics?

5.2 Purpose of the Sentence-level Readability Classifier

The purpose of this classifier is to identify sentences in the text which may be of particular difficulty for readers with autism. It is especially relevant to the task of automatic text simplification, where it could be used either ad-hoc or post-hoc. In its ad-hoc use, the classifier would help identify only those sentences which need simplification, while leaving the rest of the sen-

tences intact. In its post-hoc use, the classifier would be able to evaluate the readability of the simplified output. It is important to note that, similar to the misuse of readability formulae discussed in Section 2.3.2.1, the classifier should not be used both ad-hoc and post-hoc for the same selection of texts.

The next section presents the data used for training and evaluation of the classifier.

5.3 Corpora

The sentence-level readability classifier was trained on a set of 257 *easy* and *difficult* sentences and was evaluated on a subset of this data by using 10-fold cross-validation. The evaluation of the generalisability of the sentence-level classifier on unseen data was not feasible due to the lack of another resource with sentences with known difficulty for readers with autism.

The data used for the training and evaluation of the sentence-level readability classifier comes from two sources: the sentences from the ASD corpus (Chapter 3) (157 sentences) and sentences from Laufer and Nation's vocabulary test (Laufer & Nation 1999) (100 sentences) (Section 5.3.2). The overall training and evaluation dataset consists of 257 sentences in total, of which 125 were classified as *easy* and 132 as *difficult*.

Both datasets and their classification into *easy* and *difficult* sentences are described in detail in the subsections below.

5.3.1 Sentences from the ASD Corpus

We used the gaze data collected in order to determine the level of complexity of the sentences from the ASD corpus. Each sentence was manually defined as an area of interest and was then processed using the Gazepoint data-analysis software (Gazepoint 2015) in order to measure the number of fixations, the number of revisits and the overall reading time for each sentence.

Our initial approach to using gaze data for determining the complexity of the sentences was to divide the total dwell time for each sentence by the number of characters the sentence contained. The aim of this procedure was to obtain a raw score of reading time per character, thus normalising the data for sentence length. The reason we chose to use characters instead of words was the fact that complex sentences usually contain longer and more difficult words; hence, number of words alone is not a measure that would account for the differences in the levels of complexity. However, obtaining a raw reading-time score per character did not prove to be a suitable approach, owing to the fact that the numbers obtained for each sentence were either 0.03, 0.02 or 0.01 seconds. Because of the large number we divided by (the number of characters), longer sentences resulted in having a lower raw reading-time score per character (i.e., 0.01 seconds), while shorter sentences, which contained fewer characters, had a higher raw reading-time score per character (i.e.,0.03 seconds). This approach did not account successfully for the level of difficulty of the sentences based on their raw reading-time score,

which is why we adopted a different approach based on the literature review presented in Chapter 2.

Previous research has shown that a higher number of fixations is indicative of sentence complexity (Rayner et al. 2012), so the complexity of the sentences from the ASD corpus was eventually estimated based on the average number of fixations per sentence. The sentence classification was done by first ranking the sentences based on the average number of fixations (including revisits) per sentence and then splitting them into two classes (*easy* and *difficult* using median split [median = 10.6640]). The resulting classes contained 97 *easy* sentences from the ASD corpus and 98 *difficult* sentences from the ASD corpus.

Examples of *easy* sentences from the ASD corpus are:

"*Stretching helps loosen tight muscles and tissues.*"

"*Many animals use camouflage to blend into their surroundings.*"

"*The reef bursts with schools of tropical fish, darting among gaps in the coral.*"

"*The Spanish case provides arguments both for and against monarchy.* "

"*Whenever a new print is added, the computer compares it to all the other prints for a match.* "

Examples of *difficult* sentences from the ASD corpus are:

"*Art is always already personal and political.*"

"*The cultural gap between aristocratic royals and a more democratic populace was a major cause of Juan Carlos's fall.*"

> "*Their album "Yesterday and Today" (also known as the "Butcher Album") is highly collectible and if you have an original it is highly priced and is one of the holy grails of record collecting.*"
>
> "*Is the writing on the wall for all European royals, with their Ruritanian uniforms and gilded lifestyles?*"
>
> "*Bobby Robson was there to assess World Cup candidates, but nothing positive emerged from 90 minutes of scuffling that made one almost yearn for the more measured boredom of Rangers ' European Cup exit in Munich three days earlier.*"

The number of fixations measure is useful in measuring the cognitive load that individual sentences impose on the reader; however, it is likely that this measurement is biased towards sentence length as a main feature correlated with complexity (longer sentences require more fixations). To control for this confounding variable we balanced the training dataset by injecting an additional set of 100 short sentences with a controlled length, which were not part of the 27 texts discussed above. The next subsection describes how these additional sentences were evaluated.

5.3.2 Sentences from Laufer and Nation's Vocabulary Test

An additional set of 100 sentences from the publicly available[1] vocabulary test by Laufer and Nation (Laufer & Nation 1999) was added to the sentences from the ASD corpus.

[1]The full version of the vocabulary test containing all 100 sentences can be found here: http://www.victoria.ac.nz/lals/about/staff/paul-nation (Version A, monolingual, 20,000).

The test was initially aimed at measuring the vocabulary level of native speakers of English and consisted of 100 items containing words from five frequency levels.

Each sentence had a simple one-clause structure and all words except the target word were simple. The test used a completion item type like the following:

The story is very didactic.

a) tries hard to teach something

b) is very difficult to believe

c) deals with exciting actions

d) is written with unclear meaning

While the sentences from the ASD corpus were indicative of various linguistic phenomena which had an effect on complexity as measured by the number of fixations, the sentences from the vocabulary test included only one clause per sentence and thus their classification into *easy* and *difficult* sentences indicated purely their lexical complexity.

The sentences were evaluated by the same participants from Group 3 (Chapter 3), which consisted of 18 adults with autism (twelve male and six female) with mean age m = 37.22 (SD = 10.3), years spent in education m = 16 (SD = 3.33) and a control group of 14 adults (nine male and five female) with mean age m = 34.5 (SD = 8.19), years spent in education m = 18.93 (SD = 3.1).

The 100 sentences from the vocabulary test were classified into *easy* and

Table 5.1: Examples of *easy* and *difficult* sentences from Laufer and Nation's vocabulary test (Laufer & Nation 1999)

Easy sentences	**Difficult sentences**
It was a difficult period.	Many unwanted plants are ubiquitous.
Is this the right figure?	He treated her in a cavalier manner.
His malign influence is still felt.	She saw a bittern.
He strangled her.	The cat left a gobbet behind.
The car veered.	He rode roughshod.
This yoghurt is disgusting.	They saw the panzers getting nearer.
Look what we found in the cranny!	These thoughts obtruded themselves.
It is a marsupial.	Don't play the casuist with me!
I hate the rigmarole.	He was in a torpid state.
She loves her dachshund.	They got swingeing fines.

difficult by defining a threshold for the easy sentences (65 sentences): they had a minimum of 60% correct answers from all ASD participants. All sentences that fell under that threshold were defined as *difficult* (35 sentences).

Table 5.1 presents examples of 10 *easy* and 10 *difficult* sentences from the vocabulary test.

This section described the data used for both training and evaluation of the sentence-level readability classifier. The next section describes the features extracted for all 257 sentences.

5.4 Features

Common readability metrics such as readability formulae cannot be used for sentence-level readability (DuBay 2008), which is why we extracted a total of 48 linguistic features for each individual sentence. We used the Coh-Metrix readability assessment tool[2] for feature extraction for each individual sentence. The way each feature is derived is described in detail in McNamara et al. (2014). The features used for sentence classification were as follows.

5.4.1 Shallow Descriptors

Shallow descriptors include measures of sentence length and word length, which have been found to be strong predictors of reading ease (Dale & Chall 1948). Table 5.2 presents the Coh-Metrix 3.0 labels of the shallow features included in this study, their names and their description. In the table *"m"* stands for *mean* and *"SD"* stands for *standard deviation.*

5.4.2 Features of Cohesion

Incidence scores were computed for a number of connectives, which are indicative of different types of connections within each sentence and play an important role in the creation of cohesive links between ideas within sentences. Particular emphasis was put on the analysis of verbs which express causality and intention, as these have been found problematic for people

[2]Coh-Metrix, 3.0. Available at: http://cohmetrix.com/ [Last accessed: 7/12/16]

Table 5.2: Sentence classification: Shallow descriptors

Label	**Feature**	**Description**
DESWC	Word count	Number of words in the sentence
DESWLsy	Word length in syllables, m	Average number of syllables for all words
DESWLsyd	Word length in syllables, SD	SD of the mean number of syllables for all words
DESWLlt	Word length in letters, m	Average number of letters for all words
DESWLltd	Word length in letters, SD	SD of the mean number of letters measure
DESSL	Sentence length in words, m	Average number of words for all sentences
DESSLd	Sentence length in words, SD	SD of the mean number of words for all sentences

with autism (Martos et al. 2013). Features representing these constructs were *Causal verb incidence, Causal verbs and causal particles incidence*, and *Intentional verbs incidence*, which were all incidences of intentional actions, events, and particles per thousand words. Table 5.3 presents the features of cohesion which were included in the study, where *"inc"* stands for *incidence* and *"con"* stands for *connectives*.

5.4.3 Cognitively-motivated Features

Cognitively-motivated lexical features were also included in order to address some of the difficulties people with autism have with abstraction and unfamiliarity. These features are summarised in Table 5.4. Previous research has found that language impairments in those who have comprehension difficulties such as individuals with autism, are underlied by working memory deficits (Nation et al. 1999), hence, we included features such as *Words be-*

Table 5.3: Sentence classification: Features of cohesion

Label	**Feature**	**Description**
CNCAll	All con inc	Incidence score of all connectives
CNCCaus	Causal con inc	Incidence score of causal connectives
CNCLogic	Logical con inc	Incidence score of logical connectives
CNCADC	Adversative and contrastive con inc	Incidence score of adv. and contr. cons
CNCTemp	Temporal con inc	Incidence score of temporal connectives
CNCAdd	Additive con inc	Incidence score of additive connectives
CNCPos	Positive con inc	Incidence score of positive connectives
CNCNeg	Negative con inc	Incidence score of negative connectives
SMCAUSv	Causal verb incidence	Incidence score of causal verbs
SMCAUSvp	Causal verbs and particles incidence	Incidence score of causal verbs and part.
SMINTEp	Intentional verbs incidence	Incidence score of intentional verbs
SMCAUSlsa	LSA verb overlap	LSA overlap between verbs (Dumais 2004)
SMCAUSwn	WordNet verb overlap	WordNet overlap between verbs

Table 5.4: Sentence classification: Cognitively-motivated features

Label	**Feature**	**Description**
WRDFRQc	CELEX word freq., m	Average freq. for words in CELEX database
WRDFRQa	CELEX Log freq. (all), m	Log freq. for all words in CELEX database
WRDFRQmc	CELEX Log min freq., m	Log min. freq. for words in CELEX database
WRDAOAc	Age of acquisition, m	Age of acquisition norms from MRC
WRDFAMc	Familiarity, m	Familiarity norms from MRC
WRDCNCc	Concreteness, m	Concreteness norms from MRC
WRDIMGc	Imagability, m	Imagability norms from MRC
WRDMEAc	Meaningfulness,m	Meaningfulness norms (Nickerson & Cartwright 1984)
WRDPOLc	Polysemy, m	Number of core meanings of the word (Miller 1995)
WRDHYPn	Hypernymy for Ns, m	Sub- and superordinate WordNet relations (nouns)
WRDHYPv	Hypernymy for Vs, m	Sub- and superordinate WordNet relations (verbs)
WRDHYPnv	Hypernymy for Ns and Vs, m	WordNet relations (nouns and verbs)
SYNLE	Left embeddedness, m	Number of words before the main verb
SYNNP	Modifiers per NP, m	Number of modifiers per noun phrase

fore main verb and *Number of modifiers per noun phrase*, which account for a higher cognitive load imposed on the working memory.

All scores in Table 5.4 refer to content words unless otherwise specified.

5.4.4 Incidence Counts

A number of incidence counts were included in order to account for the syntactic density of the sentences, where the higher incidence of a feature is indicative of a higher information density within the sentence. Table 5.5

Table 5.5: Sentence classification: Incidence counts

Label	**Feature**	**Description**
DRNP	Noun phrase density, incidence	Incidence score of noun phrases
DRVP	Verb phrase density, incidence	Incidence score of verb phrases
DRAP	Adverbial phrase density, incidence	Incidence score of adverbial phrases
DRPP	Preposition phrase density, incidence	Incidence score of preposition phrases
DRPVAL	Agentless passive voice density, incidence	Incidence score of passive voice
DRNEG	Negation density, incidence	Incidence score of negations
DRGERUND	Gerund density, incidence	Incidence score of gerunds
DRINF	Infinitive density, incidence	Incidence score of infinitives
WRDNOUN	Noun incidence	Incidence score of nouns
WRDVERB	Verb incidence	Incidence score of verbs
WRDADJ	Adjective incidence	Incidence score of adjectives
WRDADV	Adverb incidence	Incidence score of adverbs
WRDPRO	Pronoun incidence	Incidence score of pronouns

presents a list of the various incidence counts used in this study.

5.5 Experimental Setup

This section presents the experimental setup for the training and evaluation of the sentence-level readability classifier.

5.5.1 Modelling Method

Similar to the modelling of the document-level classifier, the chosen modelling method for the sentence-level readability estimation is classification. This

method is selected because the main purpose of this classifier is to make a binary decision as to which sentences could benefit from automatic text simplification and which ones could remain as they are.

In the results section, we report the classification accuracy in percentages, as well as precision, recall, and F-measure.

5.5.2 Algorithms

A number of supervised learning algorithms implemented in the Weka toolkit (Frank & Witten 1998) were evaluated on the sentences. The results section below presents the best performing algorithm in terms of classification accuracy.

5.5.3 Baseline

While readability formulae are typically used as baselines for document-level readability assessment, at sentence level there is a lack of agreement on best-performing measures for sentence classification (Chapter 2). However, sentence length in words is a feature that has been present in almost all readability models and has been shown to have very high discriminatory power, which is why we selected sentence length as a baseline for our classifier. We derive the baseline accuracy by using sentence length as a single feature in the readability model.

Table 5.6: Sentence classification: Selected features

Feature	**Description**
Word count	Number of words in the sentence
Word length in syllables, m	Average number of syllables for all words
Word length in letters, m	Average number of letters for all words
Word length in letters, SD	SD of the mean number of letters measure
Intentional verbs incidence	Intentional actions, events, and particles
LSA verb overlap	Latent Semantic Analysis of verb overlap (Dumais 2004)
Pronoun incidence	Number of personal pronouns
CELEX word frequency, m	Average word frequency for all words (CELEX database)
Concreteness, m	Word concreteness measure from the MRC database (Coltheart 1981)
Imagability, m	Easiness to form a mental image of the word (MRC database)
Polysemy, m	Number of core meanings of the word (WordNet (Miller 1995))
Hypernymy for nouns, m	Sub- and superordinate WordNet relations of a target word

5.5.4 Feature Selection

The full-feature set was used to obtain a baseline model and was subsequently optimised through Best First attribute selection filter for supervised learning built in Weka (Frank & Witten 1998).

After the feature selection process, only 12 features with highest discriminative power were retained for the final model. The features are presented in Table 5.6, where *"m"* stands for *mean* and *"SD"* stands for *standard deviation.*

5.5.5 Training and Evaluation

The classifiers were trained and tested using 10-fold cross validation.

5.6 Results

Among the supervised machine-learning algorithms implemented in the Weka toolkit (Frank & Witten 1998), the best performance was achieved by the SPegasos classifier (Shalev-Shwartz et al. 2011). SPegasos, or Stochastic implementation of the "Primal Estimated sub-GrAdient SOlver for SVM" is, as suggested by its title, a type of a support vector machine (SVM) (George-Nektarios 2013). SPegasos optimises the SVM algorithm by using a "stochastic gradient descent algorithm to produce the separation hyperplane" (Shalev-Shwartz et al. 2011), or, in other words, it normalises attributes, transforms nominal attributes into binary ones and replaces all missing values (George-Nektarios 2013).

Table 5.7 presents the results for the sentence-length baseline, the baseline model including the full feature set and the final model including only the selected features. All three models were obtained using the SPegasos algorithm.

As can be seen from Table 5.7, the lowest accuracy was achieved by the model including the full feature set (74% accuracy). The sentence-length baseline is shown to be a very strong one (78% accuracy); however, it was outperformed by the selected features model, which achieved 82% accuracy.

Table 5.7: Sentence-classifier results for 10-fold cross-validation

	Baseline	All features	Selected features
Precision	0.816	0.745	0.841
Recall	0.79	0.743	0.817
F	0.787	0.743	0.815
Accuracy	0.78	0.74	0.82

5.7 Discussion

This section presents a discussion of the main findings and methodological challenges from the development of the sentence-level readability classifier.

5.7.1 Methodological Challenges and Contributions

The selected-features classifier presented in this chapter had a satisfactory accuracy (82%) compared to other sentence-level classifiers such as the READ-IT classifier (Dell'Orletta et al. 2011), which achieved 78.2% accuracy for sentences in Italian; the classifier by (Vajjala & Meurers 2014) where 80% accuracy was achieved by using pairs of original and manually simplified sentences from news articles; and the one by (Pilán et al. 2014) where 71% accuracy was achieved when classifying Swedish sentences for foreign language-learners.

In addition to the comparatively good accuracy achieved, this is the first sentence-level classifier to use a gold standard based on gaze fixations and comprehension testing of particular sentences. By comparison, all studies

mentioned above rely on manual simplification by experts or on sentences obtained from simple and complex texts as a gold standard of text accessibility. Furthermore, all sentences in our training and test data are naturally occurring sentences as opposed to simplified versions of other sentences, where sentence length and lexical complexity have been manipulated and thus may have introduced bias. Finally, the complexity of the sentences from the ASD corpus is evaluated as they appeared in the context of a coherent text.

However, in spite of the fact that the selected-features classifier outperformed the other two classifiers, we observed that the baseline of sentence length alone is a very strong one, as our classifier was only 4% more accurate (the feature of sentence length alone achieved approximately 78% accuracy compared to the selected-feature set which achieved approximately 82%). This very strong result for the baseline was achieved even though our dataset featured a hundred one-clause sentences, reducing a potential bias towards classifying longer sentences as more complex ones due to the gaze fixation measure. In this context, a bias towards sentence length means that more false negatives may have been introduced by the gaze fixations measure where difficult short sentences could have been mistakenly classified as *easy*. Admittedly, however, a large body of research has confirmed that long sentences impose a heavier cognitive load on the reader (Chapter 2) and thus a classifier would not be wrong in learning that they are more complex.

5.7.2 Limitations

The use of gaze data as a gold standard for complexity inevitably introduces implications resulting from eye-tracking inaccuracies. In spite of the robust measures taken to reduce noise in the data (Chapter 3), it is possible that some of the gaze fixations may have been shifted. However, this type of inaccuracy is likely to have a less dramatic effect at sentence level (compared to word level, for example), where, even if a fixation is shifted it will still likely fall into the same sentence. Introducing buttons as anchors for the adjustment of the gaze path (Chapter 3) allowed us to detect and regulate these shifts as much as possible. Discarding noisy or inaccurate data led to another limitation of the sentence-level classifier, namely the small number of participants involved in the assessment of the sentences. This limitation has been compensated for by the fact that using eye tracking produces a large number of data points which could, to a certain extent, compensate for the small number of readers.

5.8 Summary

This chapter has presented the dataset, features and modelling method used in the development of a sentence-level readability classifier. It presented a new gold standard of sentence readability, which was based on gaze data and comprehension testing as opposed to simplification by experts. The classification accuracy achieved by the classifier was comparable to the accuracy

for sentence readability reported in previous work. The classifier also outperformed the baseline of sentence length; however, the difference in accuracy was only 4%, suggesting that sentence length alone could be used as a strong predictor of sentence readability.

This chapter and the previous two have presented work concerning readability assessment for readers with autism. The next chapters will investigate the effects images have on comprehension and memorisation (Chapter 6) and the way web users with autism search for information on web pages (Chapter 7).

Chapter 6

Images in Text: Effects on Comprehension, Memorisation and Attention in Readers with Autism

6.1 Chapter Overview

In previous chapters we discussed ways in which to measure the readability of text content for readers with autism. However, the comprehension and memorisation of information obtained through reading does not depend solely on the linguistic characteristics of a text. People with autism are known to be very strong visual thinkers and to have an above-average ability to process visual information (Kana et al. 2006, Grandin 2009, Quill 1997, Dettmer et al. 2000), which is why we investigate whether the insertion of images into text could be used to enhance their reading experience by using their strengths to compensate for their difficulties.

In this chapter we will discuss a series of studies into the effect of images inserted into text on comprehension, memorisation and attention in readers with autism, as a way to improve their reading performance. The results of these experiments will be synthesised in a set of accessibility guidelines and are aimed at addressing the issue of insufficient instruction with regard to

images in the manuals for easy-to-read documents. Furthermore, the results of these studies have implications for the development of web content.

- The first study presented in this chapter investigates between-group differences in the proportion of time spent looking at an image or a text paragraph for 39 image and text pairs based on eye-tracking measures. Insights into the way attention shifts between visual and linguistic stimuli in readers with autism could help improve user-centred document layouts and software interfaces.

- In addition to attention shifting, we investigate the effects of image type (i.e., photograph or symbol) on attention in adults with high-functioning autism. Distinguishing between images with high resemblance to their referent in reality (photographs) and images with low resemblance to their referent in reality (symbols), as well as investigation of their effect on attention is important for readers with autism due to their difficulty processing vague visual representations (Hartley & Allen 2014, Allen 2009). Currently both types of images are used in easy-to-read documents and in language-assistance tools for people with autism.

- Following these experiments, reading tests were conducted to investigate the effects of images included in text on the reading comprehension of adults with autism, as well as on their memorisation and recall of information. These effects were investigated in both within-groups and

between-groups.

- Finally, user preferences and the perceived level of difficulty of easy-to-read texts in adults with high-functioning autism were investigated in order to establish whether easy-to-read documents are perceived as too difficult or too easy, thus potentially leading to loss of interest and reduced concentration.

The experiments in this chapter address research question four:

RQ4: Do images inserted into texts have an effect on participants' attention, comprehension and memorisation of a text, measured both objectively and subjectively?

The insights gained from the experiments and their summarisation into accessibility guidelines are the fourth original contribution of the thesis:

Contribution 4. Improved text- and web-accessibility guidelines for people with autism.

Some of the experiments and results presented in this chapter were presented in Yaneva et al. (2015) and in Yaneva et al. (n.d.) (under review). Some of the characteristics of easy-to-read documents were analysed and presented in Yaneva (2015).

The next section presents the motivation behind the experiments.

6.2 Motivation

This section presents the motivation behind the choice of investigating images as a particular type of a comprehension cue for readers with autism.

6.2.1 Images in Text Documents and Assistive Software for People with Autism

Currently, images are widely used both in easy-to-read documents (Tronbacke 1997, Freyhoff et al. 1998) and in language assistance tools developed for people with autism, such as *Puzzle Spelling Words*[1], *VAST-Autism*[2], *Stories About Me*[3], *OpenBook*[4], etc.

However, there is no information regarding guidelines for images which have been used when developing these applications and little is known about the way images should be used in assistive software. There are few existing user-requirement surveys conducted specifically for people with autism and some of them touch upon the issue of images only very slightly; one study stresses that no bright colours or background images should be used (Pavlov

[1]Puzzle Spelling Words. Available at: http://touchautism.com/app/puzzle-spelling-words/, [Last accessed: 22/03/2016]

[2]VAST-Autism. Available at: http://www.speakinmotion.com/solutions/mobile-apps/vast-autism-series/, [Last accessed: 22/03/2016]

[3]Stories About Me. Available at: https://itunes.apple.com/us/app/stories-about-me/id531603747?mt=8, [Last accessed: 22/03/2016]

[4]OpenBook (software). Available at: http://www.openbooktool.net/, [Last accessed: April, 2015]

2014), while a survey with 120 autistic respondents and their families concludes that sensory integration and attention issues should be addressed by allowing users to set colours or sounds (Putnam & Chong 2008). None of these studies investigated preferences on the use of images or visual cues in software. However, one study highlighted "the issue of identification of the most appropriate set of pictures for this system" (Sampath et al. 2010) and continued with the following recommendation:

> " *In [the] case of a child with autism, due to their difficulty with abstraction and generalization, the pictures need to have a strong resemblance to their referents. The more relevant these pictures are to the child's culture and environment, the easier it is for them to use the system. (Sampath et al. 2010, p. 35)*"

This statement was not empirically tested; however, it raises an important question since, currently, easy-to-read documents and assistive software for readers with autism use miscellaneous types of images, owing to lack of guidelines regarding image type.

6.2.2 Symbolic Understanding of Images in People with Autism

The Oxford Dictionary of English defines "symbol" as "a mark or character used as a conventional representation of an object, function, or process" [5]. The cognitive processing of symbols and photographs, as, respectively, weak and strong representations of their referred objects, requires two different

[5]Oxford English Dictionary [Online] Available at: http://dictionary.oed.com]

levels of symbolic understanding. In the process of childhood development, children first play with real objects and learn to associate them with the activity they are used for. When they learn to identify the real object with a photograph of it, they demonstrate a higher level of symbolic understanding, the next step of which is learning to match the object to a drawing or a symbol and which culminates in the acquisition of a word to denote this whole set of entities (DeLoache 2008, Bialystok 2000). In the case of easy-to-read documents, both images with a strong resemblance to their intended referent (photographs) and images with a low resemblance (drawings and symbols) have been widely used (Figure 6.1).

Figure 6.1: Examples of a symbol and text pair and a photograph and text pair

Without wishing to stray into the philosophical debate around the differences between symbols and signs, this thesis uses the term "symbol" to refer to images in adapted documents rather than to photographs of real people or objects, for example.

Children with autism are considered to have greater difficulty decoding

vague representations, compared to typically developing children, owing to their impaired ability to generalise, grasp context, or reason about the intent of the author (Hartley & Allen 2014, Allen 2009). This evidence suggests that the type of images used would have a greater impact on the perception of autistic users than on that of the neurotypical (non-autistic) ones. While vague representations are found challenging by children with autism, there have been no studies, to the best of our knowledge, investigating the development of symbolic understanding in *adults* with autism.

In addition to potential difficulties with symbolic understanding, there are also differences in attention patterns between people with and without autism. These differences are so prominent that their first mention dates back to as early as the first mention of autism by Leo Kanner in 1943 (Kanner 1943). Atypical attention patterns may have implications for reading, as readers with autism are thought to focus on fragments of information rather than perceiving the text as a whole with various relationships connecting the fragments together (Happé & Frith 2006).

Currently, there is no understanding as to whether readers with autism process images in text differently from readers without autism. Finally, because of attention differences between the two populations, if readers with autism focus longer on images in text compared to readers without autism, it is uncertain whether they do this because images provide them with comprehension cues or because images distract them.

In the sections below we present a series of experiments aiming to bring

clarity to these issues and to gain an understanding as to the best way to use images to support readers with autism.

6.3 Study Hypotheses

This section presents the hypotheses tested in this chapter. The research consists of initial-stage experiments (hypotheses 1-4) initially presented in Yaneva et al. (2015) and henceforth referred to as Study 1; and follow-up experiments (hypotheses 5 - 12), henceforth referred to as Study 2.

The participants in Study 1 were 20 people diagnosed with autism and 20 non-autistic control subjects, who all read nine texts, while having their eye movements recorded by an eye tracker. This study investigated questions relating to attention, as measured by gaze-fixation data and ease of comprehension, which, at this stage, was only measured subjectively through the use of Likert scales.

The study investigated whether there were any between-group differences in the proportion of time each group spends looking at the image in 39 text and image pairs (see Figure 6.1). The study also investigated which type of images, photographs or symbols, elicited longer fixation times and thus impose a heavier cognitive load on the participants. We also wanted to find out how the level of difficulty of easy-to-read documents would be perceived by adults on the spectrum. Comparing the perceived level of difficulty of the nine texts presented also gave information on whether texts considered

to be written in Plain English but with different readability levels evoked different responses from the participants. Finally, we investigated what the text-presentation preferences of the two groups were, by including a survey question at the end of the experiment. The follow-up study featured several additional user-preference questions, which gave us a more in-depth understanding of the preference differences between autistic and non-autistic users. These research questions are summarised in the following 4 hypotheses:

H.1: There is no difference between groups in the proportion of time spent looking at the image for each text and image pair.

H.2: Within groups, there is no effect on the time spent looking at photographs and symbols.

H.3: There is no difference between groups in the perceived level of difficulty of the presented documents.

H.4: There is no difference between groups regarding the text presentation preferences.

The design and procedure of the experiment testing these hypotheses are presented in Section 6.4.

After testing these hypotheses, a follow-up study with 18 autistic and 18 non-autistic participants was conducted in order to find out whether the

increased attention paid to images by the autistic participants (Section 6.5) was related to their use of images as comprehension and memorisation cues, or, on the contrary, whether the images prevented them from concentrating on the text, or had no effect at all. Comprehension and memorisation were measured through objective measures such as literal and inferential multiple-choice questions, which were asked either immediately after the text had been read (comprehension) or approximately five minutes later, after all texts had been read and survey questions had been answered (memorisation). Comprehension and memorisation were also measured by subjective measures, i.e., the answers given to two survey questions: *"Do you think the insertion of images into some of the texts helped you comprehend the text better?"* and *"Do you think the insertion of images into some of the texts helped you memorise the information better?"*. In the survey following the experiment, the subjective usefulness of images was measured again through asking another question, worded differently from the first, in order to ensure that the answers given were not dependent on the way the questions were phrased: *"Which did you find most useful: a) reading texts* ***with*** *images, b) reading texts* ***without*** *images, c) no preference"* (Section 6.4). We tested the effects of images on comprehension and memorisation both within groups and between groups, allowing investigation into autism-specific patterns of reading behaviour. The formal hypotheses relating to the questions from the follow-up study are presented below, where objective measures are answers to multiple-choice comprehension questions and subjective measures are the

answers participants gave to survey questions.

We first present results from within-group comparison investigating the following hypotheses:

H.5: Within groups, images included in the text do not have an effect on **comprehension** as measured through **objective measures**.

H.6: Within groups, images included in the text do not have an effect on **memorisation** as measured through **objective measures**.

H.7: Within groups, images included in the text do not have an effect on **comprehension** as measured through **subjective measures**.

H.8: Within groups, images included in the text do not have an effect on **memorisation** as measured through **subjective measures**.

We then present results from between-group comparisons defined by the following hypotheses:

H.9: Between groups, images included in the text do not have an effect on **comprehension** as measured through **objective measures**.

H.10: Between groups, images included in the text do not have an effect

on **memorisation** as measured through **objective measures**.

H.11: Between groups, images included in the text do not have an effect on **comprehension** as measured through **subjective measures**.

H.12: Between groups, images included in the text do not have an effect on **memorisation** as measured through **subjective measures**.

The next section presents a detailed description of the experimental design and the measures used in this study.

6.4 Method

This section describes the experimental design of the experiments testing the hypotheses listed above. The first study investigates differences in attention and user preferences (Yaneva et al. 2015), while the second study investigates the effects of images on comprehension and memorisation. Both studies implemented between-group and within-group comparison designs. Study 1 was conducted to test hypotheses 1-4 (presented in the previous section) and Study 2 tested hypotheses 5-12, while also helping to test hypothesis 4.

6.4.1 Design

6.4.1.1 Study 1

The study implemented both between-group and within-group comparison design, where the independent variable was the use of images in texts, and had three levels: texts with photographs (20 photographs in total), texts with symbols (19 symbols in total) and plain texts (with no images). After reading each text and having their reading fixations recorded by an eye-tracking device, the participants were asked one literal multiple-choice question about the meaning of the text, in order to ensure that they were reading for meaning as opposed to just skimming through the text. The questions testing the comprehension of the participants were chosen to be literal owing to the simplicity of the easy-to-read texts, where, by default, no strong inferential or reorganisational skills are needed in order to comprehend their meaning. In-depth reading comprehension was later tested in Study 2, which involved various types of inferential questions.

As an example, a text about eating habits, where various types of foods were discussed, would be followed by a literal multiple-choice question with only two possible answers:

High-fibre foods include: a) Meat and milk or b) Bread and beans?

Knowing that they would need to answer a question after the text's removal from sight, all participants read the documents carefully (as evidenced by the gaze-pattern videos produced by the eye tracker) and were all able to

answer 100% of the questions correctly. The answers to these questions are used as a control variable only and are not included in the analysis of this study. After reading each text and answering the multiple-choice questions, participants would rate their subjective perception of the difficulty of the text on a Likert scale from 1 to 5, where 1 stood for "very easy" and 5 stood for "very difficult". Finally, all participants answered a question about their reading preference, where they could choose between reading texts with a) photographs, b) symbols, or c) plain text (no images) or choose d) "It makes no difference to me". Participants were allowed to choose none or more than one of the answers and were encouraged to elaborate on their choice if they wanted to. Based on the above design, we considered six metrics in total. Images and text paragraphs were defined as areas of interest (AOIs) and a number of gaze-based metrics were obtained based on how many times and for how long participants looked at these areas. The metrics used in this study are:

Average Time Viewed (ATV): The average time an AOI was viewed by all participants measured in seconds. This is an average from the total dwell time, including the durations of all fixations and all revisits.

Average Number of Fixations (AF): The average number of gaze fixations from all participants in a given AOI.

Average Number of Revisits (AR): The average number of go-back gaze fixations from all participants in a given AOI. Go-back gaze fixations are fixations in the span of a given AOI elicited after the gaze path has left

the AOI and has then returned to revisit it. Revisits are a valuable source of information for heavy cognitive load and the way information from different parts of the screen is integrated.

Reading-time score: This measure was developed by estimating the mean reading time per text in each group and then dividing the result by the number of words in the text. This was done in order to control for the differences in length between the nine texts. Reading time has been used as an indicator of reading difficulty, with examples of texts similar in length but differing in the time they require for reading based on their complexity level DuBay (2008).

Perceived level of difficulty: This measure was obtained through the Likert scale results reported by participants after each text. This measure was chosen instead of reading-comprehension questions as it more accurately reflects the subjective impressions of the participants regarding text difficulty and is thus more useful for evaluating their attitudes towards the difficulty of Plain English texts.

Text-presentation preference: Information was gathered through the following survey question: "In your everyday life, do you prefer reading texts with: a) photographs, illustrating the main ideas b) symbols, illustrating the main ideas, c) plain texts without any images or d) "It makes no difference to me".

6.4.1.2 Study 2

Study 2 was conducted approximately eight months after Study 1 with the aim of addressing and following up on a number of interesting questions posed by the results of Study 1, presented in Section 6.5. The main focus of Study 2 was to establish whether the insertion of images in text has an effect on comprehension and memorisation of the information being read and to gain a deeper and broader understanding of the user preferences of people with autism with regard to the use of visual cues.

Study 2 involved three educational texts with fairly high levels of complexity, which were modified to satisfy two conditions: text-only documents and documents where complex words in the texts were illustrated by images. Each participant saw only one version of the document and was presented both with texts which included images and other texts which did not include images. For each text there were 4 complex words or phrases illustrated by images, as shown in Figures 2 and 3. The words and phrases illustrated by images were:

Cytoplasm, Nucleus, Mitochondria, Vacuole (Text 1);

Galaxies, Milky Way, Satellites, Comets (Text 2)

Electronic circuits, Breadboard, Programmable Interface Controllers (PICs), Flowchart (Text 3).

The criteria for selecting words to be illustrated was based on their complexity as measured through word length and word frequency, assuming that

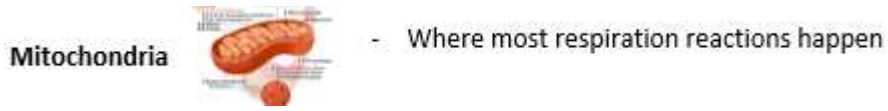

Figure 6.2: Example of an illustrated complex word (Text 1, Study 2)

Figure 6.3: Example of illustrated complex words or phrases (Text 3, Study 2)

longer and less frequent words are more difficult to understand (DuBay 2008, Pastor et al. 2008). We limited this study to nouns and noun phrases, which in our educational texts were used as terms. This was done because, for the purposes of the experiment, the words had to be concrete and depicted unambiguously, which is not a straightforward task for other domains where words tend to be polysemous or abstract.

The final selection of images was made by a human (the author), after the three texts were automatically processed to retrieve up to ten images associated with words and phrases from the texts. Two sources of images were used: ImageNet (Fei-Fei & Russakovsky 2013) and Wikipedia[6]. ImageNet is a database aligned to the WordNet noun hierarchy in which humans

[6]Available at: https://en.wikipedia.org/wiki/MainPage

manually assigned images to each node (Miller et al. 1990). The automatic processing identified nouns from the texts which are present in ImageNet and retrieved the first ten images associated with them. ImageNet proved to be fairly incomplete for the purposes of our task. For this reason, Wikipedia was also used to retrieve images from Wikipedia articles related to phrases from text.

Comprehension and memorisation were measured through multiple-choice questions, while the subjective perception of the effects of images on comprehension and memorisation was measured through survey questions. An additional survey question was used to gain information on user preferences towards the positioning of images.

Objective measures of comprehension: The level of comprehension of each participant for each text was measured through two multiple-choice questions per text with three possible answers per question. Participants were allowed to re-read the text as many times as they needed but were not allowed to look at it while answering the questions. The questions examined three types of reading comprehension starting with literal understanding, re-organisation skills (where the participants are required to combine explicit information from different parts of the text in order to obtain a third piece of information) and finally, the ability to make gap inferences (where the reader is required to combine two pieces of implicit information from the text in order arrive at a third piece of information, which is also implicit) (Day & Park 2005). These types of comprehension questions were chosen because autistic

readers have previously been shown to have difficulties making use of context to answer reorganisation and gap-inference questions (Saldaña & Frith 2007, O'Connor & Klein 2004).

Subjective measures of comprehension: The subjective perception of the participants of the effects images had on their comprehension was measured through the following survey question: *"Do you think that the inclusion of images in some of the texts helped you understand them better?"* with possible answers *a) Yes, b) No and c) Cannot say.* To ensure that the answers of the participants were not influenced by the phrasing of the question, a similar question was asked later on in the survey: *"Which did you find most useful: a) Reading texts with images, b) Reading texts without images, c) I have no preference"*, where the concept of comprehension was substituted by the more general concept of "usefulness".

Objective measures of memorisation: Memorisation of the important information from the texts was measured through two multiple-choice questions with three possible answers each, which were asked after the participant had read all three texts and answered all comprehension questions and all survey questions. Participants were not allowed to look at the texts while answering the memorisation questions. There were roughly five minutes on average between the reading of the texts and the answering of the memory questions, during which time participants' concentration on the meaning of the particular texts was interrupted by the survey questions they were asked.

In a similar way to the comprehension questions described above, the memorisation questions examined literal understanding, reorganisation skills and the ability to make gap inferences.

Subjective measures of memorisation: Similar to the subjective measurement of comprehension described above, subjective perception of the effects of images on memorisation was measured using a survey question: *"Do you think that the inclusion of images in some of the texts helped you memorise the information better?", with answers a) Yes, b) No and c) Cannot say.* The general question about the usefulness of images which was asked in Study 1, is also relevant to the memorisation of the information: *"Which did you find most useful: a) Reading texts with images, b) Reading texts without images, c) I have no preference".* Finally, the answers to all of the survey questions on the subjective perception of images are compared to the answers to the survey question from Study 1 about preferences concerning reading texts with and without images.

Preferences to image positioning: In order to determine what the best positioning of images in text would be, we designed three versions of the same text document (Text 1). in each version, images were positioned in a different place: either above the word, on the right-hand side of the word or on the right-hand side of the line, i.e., at the end of the sentence in which the target word was included. We asked the participants to answer the following question: *"Which positioning of the images do you prefer? a) Document 1:*

images on the right-hand side of the word, b) Document 2: images on the right-hand side of the sentence, c) Document 3: images above the word, d) Other, please specify".

6.4.2 Participants

Study 1 included 20 adults (seven female, 13 male) with a confirmed diagnosis of autism (n=10 Autism Spectrum Disorder, n = 9 Asperger's syndrome and n = 1 semantic-pragmatic disorder), who were recruited through four local charity organisations. The control group comprised 20 non-autistic adults (11 female and nine male). None of the 40 participants had comorbid conditions affecting their reading (e.g. dyslexia, learning difficulties, aphasia etc.), but some participants were diagnosed with comorbid depression (n = 4, ASD group; n = 1, control group) and anxiety (n = 6, ASD group). Mean age (m) for the ASD group was m = 30.75, with standard deviation SD = 8.23, while for the control group mean age was m = 30.81, SD = 4.8. Mean number of years spent in education, as a factor influencing reading skills, for the ASD group was m = 15.31, SD = 2.9, and for the control group, m = 17.25, SD = 2.15. None of the participants in the two groups were diagnosed with a learning disability or a developmental delay, so no matching for mental age was required (Jarrold & Brock 2004). All participants were native speakers of English and had normal or corrected vision. Results from three participants from the ASD group were discarded due to poor calibration or data loss (i.e., too many head movements during reading), resulting in dramatic

inaccuracies in more than 70% of the data collected from them. Hence, the results analysed were obtained from 17 ASD and 20 control participants.

The inclusion and exclusion criteria for the participants in Study 2 were the same as in Study 1. Study 2 involved 18 adults with a confirmed diagnosis of autism (11 male and seven female) and 18 control-group neurotypical (non-autistic) participants (12 male and six female). Of the autistic participants in Study 2, 12 had taken part in Study 1 and six were newly-recruited; hence, there was a total of 26 different ASD participants in these studies. Mean age in years for the ASD group was m = 36.83, with standard deviation SD = 10.8 and mean number of years spent in education as a factor influencing reading skills was m = 16, SD = 3.33. None of the participants had been diagnosed with a learning disability, dyslexia or with developmental delay, which, as in Study 1, meant that participants did not need to be matched based on their mental age (Jarrold & Brock 2004). As in Study 1, all participants were native speakers of English.

6.4.3 Materials

Both Study 1 and Study 2 were designed with the specific characteristics of autism in mind, which ruled out the inclusion of a large number of documents for assessment. This limitation was due to the fact that people with autism have been shown to have difficulties concentrating for long periods of time (Brugha et al. 2012, Sasson & Elison 2012), which prevents them from reading many texts. They also tend to need longer to comprehend instructions,

to calibrate an eye-tracker (Quill 1997) and to accustom themselves to an unfamiliar laboratory environment. These all contributed to higher levels of social anxiety among them (Bejerot & Mörtberg 2014).

6.4.3.1 Study 1

The materials used in Study 1 were easy-to-read documents. In order to ensure that the texts included in this study were representative of the easy-to-read information available to people with special needs, they were selected from a pool of 100 easy-to-read documents: from this pool, a sample of seven texts comprising 39 image and text snippets was carefully chosen for the experiment (Table 1), based on the following criteria:

Topic (none of the documents included required any prior knowledge, nor did any of them discuss sensitive topics),

Source (the selected documents came from all of the sources listed above, such as charity organisations, government and healthcare departments),

Readability level (documents, or parts of documents, were included so as to cover a diverse range of readability levels), and

Images (both photographs [n = 20] and symbols [n = 19], were included, each image accompanied by paragraphs of text as opposed to one- or two-word descriptions).

As the easy-to-read documents contain images by default, the two texts classified as "plain text without images" were selected from the WeeBit readability corpus (Vajjala & Meurers 2012). They were written according to

Table 6.1: Characteristics of the texts included in Study 1

	Type	**Words**	**Images**	**FKGL**	**Flesch**
T1	Photos	77	4	8.16	60.11
T2	Photos	96	5	6.73	67.33
T3	Symbols	74	6	2.71	92.54
T4	Photos	178	8	5.52	75.33
T5	Symbols	77	6	5.79	70.67
T6	Symbols	121	6	1.75	95.00
T7	Photos	58	4	6.63	68.16
T8	None	178	0	4.67	80.22
T9	None	163	0	4.93	79.548

Plain English requirements and their readability scores were medium compared to the seven easy-to-read documents selected. Thus, the study included nine texts overall, the details of which are summarised in Table 6.1, where "FKGL" stands for "Flesch-Kincaid Grade Level" (Kincaid et al. 1975) and "Flesch" stands for "Flesch Reading Ease" readability formula (Flesch 1948). Readability scores were obtained using the Coh-Metrix 3.0 software (McNamara et al. 2010).

6.4.3.2 Study 2

The materials in Study 2 were three educational texts from the domains of Biology, Astronomy and Computer Science (Table 6.2). The educational do-

Table 6.2: Characteristics of the texts included in Study 2

	Domain	**Words**	**FKGL**	**Flesch**
T1	Educational	101	8.229	55.011
T2	Educational	100	2.943	94.15
T3	Educational	113	6.963	67.304

main was chosen for the reasons that higher-level educational texts are not too easy to understand, that increasing people with autism's comprehension of such texts is expected to have the greatest impact on their lives and, finally, because educational texts contain many complex words or terms. Furthermore, in such texts, terms are used to represent a specific concept, meaning that they are not polysemous and are thus more suitable for illustration by an automatic illustration system. These three texts were significantly more complex than the texts in Study 1 and featured material corresponding to the GCSE level of the British educational system (16 years of age). All three texts were obtained from the WeeBit corpus (Vajjala & Meurers 2012) and were chosen in such a way that they would not require prior knowledge in order to be comprehended. For example, the text from the domain of Biology was entitled "Cells" and introduced information about what cells are, what types of cells there are in animals and plants and what the components of plant and animal cells are.

Information about the choice of particular words being illustrated, as well as the way images were obtained, is described in Section 6.4.1.

6.4.4 Apparatus

The device used for recording the gaze of the participants during task performance in Study 1 (Yaneva et al. 2015) was a Gazepoint GP3 video-based eye tracker (Gazepoint 2015) (60Hz sampling rate), with a 19" LCD monitor. No equipment was attached to the heads of the participants or anywhere else on their bodies. The eye tracker was calibrated individually for each participant using a nine-point calibration procedure. The use of a chin rest is not recommended due to sensory issues common within autism (Sasson & Elison 2012), which is why the distance between each participant and the eye tracker was controlled by using a fixed chair only, and was roughly 85 cm.

6.4.5 Procedure

For both Studies 1 and 2, each participant was given verbal instructions on the sequential order in which the experiment would proceed and on how the eye tracker would function (for Study 1). Each participant was given the opportunity to ask questions and to request a break at any point if he or she felt tired. The eye tracker was recalibrated if the participants needed to get up during their breaks. Demographic information (age, education, diagnosis) was collected after the instructions were given. After that, the instructions were repeated and the calibration procedure was started. Each of the documents was presented on-screen in a randomised order and participants could take as long as they needed to read it. After reading each document, the participant would be asked a comprehension question verbally, without hav-

ing the opportunity to look at the text while answering. At the end of the experiment the survey question was asked and participants were debriefed. In Study 2, after the survey questions were answered the participants had to answer memorisation questions, as described in Section 6.4.1.

6.5 Results

This section presents the results of the two studies. The first three subsections present results from Study 1 (Yaneva et al. 2015), while the rest of the section presents results from Study 2. The last subsection discusses the formal testing of Hypothesis 4, including data from both Study 1 and 2, which is why it is featured towards the end of the section.

Fixation points from the eye tracker and their grouping into specific areas of interest (AOIs) such as images or text paragraphs, were analysed using the Gazepoint analysis system, specifically developed for the GP3 Gazepoint eye trackers (Gazepoint 2015). Statistical data were analysed using the IBM SPSS Statistics software, Version 20 (IBM Corp. 2011).

6.5.1 Attention to Images

In this subsection we test hypothesis one, which stated:

H.1: *There is no difference* ***between groups*** *in the proportion of time spent looking at the image for each text and image pair.*

First, we compared the overall time participants from both groups spent

looking at the 39 images. A Shapiro-Wilk test showed that all of the gaze-based measures, namely Average Time Viewed (ATV), Average Fixations (AF) and Average Revisits (AR), were non-normally distributed. Hence, the study used a Mann-Whitney U test to assess the null hypothesis using all three gaze-based measures. The test clearly rejected this hypothesis, confirming a difference between the groups, where the participants with autism were shown to spend significantly more time not only looking at the images, (ATV: U=338.5, N1=17, N2=20, p=0.000, two-tailed; AF: U= 290.000, N1=17, N2=20, p=0.000,two-tailed) but also revisiting them (AR: U=331.000, N =17, N2=20, p=0.000, two-tailed).

However, a significant difference between the absolute average viewing times participants spent looking at the images may have resulted from longer overall reading times in the ASD group. To investigate this further, for each group we added up the Average Viewing Times (ATV) for each image together with the ATV of its corresponding paragraph resulting in an "ATV-total" measure for each text-image pair, representing 100% of the time spent looking at the text and image together. We thus had AOIs containing the ATV-total for 39 text-image pairs. Then the ATV of each image was computed as a percentage of the ATV-total for each pair in the following way: ATV per image (%) = 100% — ATV per text paragraph (%). A Mann-Whitney U test indicated that there was a statistically significant difference between the two proportions (U=461.00, N =17, N2=20, p= 0.003, two-tailed), and thus **H.1** was rejected, with the ASD group spending a greater

proportion of time on images compared to the control group, which is evidence of an atypical attention pattern in this population. The proportion of time the ASD group spent looking at the images totalled 20.32%, compared with 13.42% for the control group, leaving the ASD group with 79.68% of their time spent on reading the text and 86.58% for the control group.

6.5.2 Photographs versus Symbols

H.2: *Within groups, there is no effect on the time spent looking at photographs and symbols.*

A Shapiro-Wilk test showed that the data were non-normally distributed for Average Time Viewed (ATV) (Symbols: $p = 0.001$, Photographs: p=0.011) and Average Revisits (AR) (Symbols: p=0.000, Photographs: p=0.163) in the control group, while the Average Fixations (AF) dataset for both the control and ASD groups and the datasets ATV and AR for the ASD group were normally distributed (Control group: Symbols p=0.001, Photographs p=0.011; ASD group: Symbols p=0.091, Photographs p=0.332). Hence, a paired-samples t-test was used to compare the data in the "symbol" and "photograph" classes for the ASD group for all three measures and for the AF dataset from the control group. A Wilcoxon Matched Pairs test was in turn used to compare the non-normally distributed ATV and AR datasets for the control group. First, Tukey's test for outliers was carried out, which showed that there were no outlier values in the datasets. The paired-samples

t-test showed that in the ASD group there was no significant difference between the time spent viewing images according to the ATV measure (t = -1.389, df =18, p = 0.182, two-tailed), AF (t = -1.339, df = 18, p = 0.197, two-tailed) or AR (t = 0.378, df = 17, p = 0.710, two-tailed). Similarly, the results from the Wilcoxon Matched Pairs test revealed no significant difference between the times participants in the control group spent looking at symbols or photographs for the ATV and AR measures (ATV measure: z = -0.765, N—Ties = 19, p=0.444, one-tailed; AR measure: z = -0.763, N—Ties = 17, p=0.445, one-tailed), and a paired-samples t-test confirmed the same for the AF measure (t = -0.298, df = 18, p = 0.769). The results failed to reject the **H.2** hypothesis that there is no significant difference between the times participants from the two different groups look at photographs and symbols. The results indicate that photographs and symbols impose similar cognitive loads on the participants from the two groups and thus both sets are equally suitable for use in easy-to-read documents for adults with ASD.

6.5.3 Level of Difficulty

H.3: *There is no difference* ***between groups*** *in the perceived level of difficulty of the presented documents.*

A Shapiro-Wilk test showed that for the ASD group the data were not normally distributed for all texts, with the exception of Text 8, and for the control group the data were not normally distributed for all texts. Hence,

to study the occurrence of any significant differences between the perceived level of difficulty in the nine texts, we used a Friedman's non-parametric test for repeated measures, which showed no statistically significant difference between the perceived level of difficulty in both groups for all nine texts (ASD group: $\chi^2(8) = 9.679$, $p = 0.139$, control group: $\chi^2(8) = 10.145$, $p = 0.119$), indicating that documents with readability levels between 61 and 95 Flesch Reading Ease score are considered as the same class of difficulty by the ASD group.

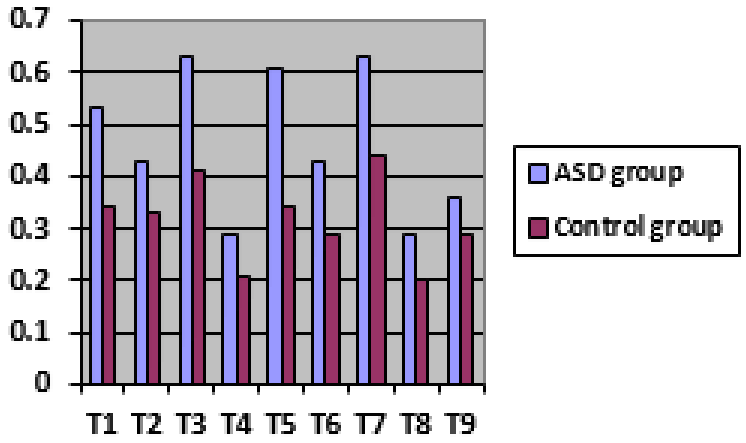

Figure 6.4: Differences in reading time scores between the autistic and non-autistic participants

Nevertheless, there were expected between-group differences in the reading time for each document, showing that despite the lack of developmental delay, the ASD group did struggle more when reading the nine texts (Figure 6.4). Furthermore, the ASD group rated the perceived level of difficulty of the texts less consistently, with answers ranging between very easy (n =

54), easy (n = 37) and medium (n = 23) to even reaching difficult (n = 4) and very difficult (n = 2). The control group, on the other hand, tended to answer very easy (n = 117) and easy (n = 20) and none of the participants ranked any text as difficult or very difficult. There was a statistically significant difference between the perceived level of difficulty for the two groups, as measured using the Mann Whitney U test (U = 4952; $p = 0.000$).

6.5.4 Effects of Images on Text Comprehension

***H.5**: **Within groups**, images included in the text do not have an effect on **comprehension** as measured through **objective measures**.*

In order to test whether images have an effect on comprehension we first compared the answers to the immediate multiple-choice questions (objective measures) for two conditions: one, where a text was presented with images, and another one, where the same text was presented without images. The results from a Mann-Whitney U test indicated that there was no within-group effect for the ASD group (U = 1377, p = 0.539, two-tailed), nor for the control group of non-autistic participants (U = 1431, p = 0.801, two-tailed).

In studies with small sample sizes like the one in this experiment, a lack of a statistically significant effect could be attributed to an insufficient number of participants. To make sure that in the case of this study the lack of an effect was not due to the small size of the participant sample, we measured

the size of the effect by using Cohen's r formula by Cohen (1988), which has been recommended as the best effect size measure for studies involving non-parametric data and for studies using the Mann-Whitney U test in particular (Fritz et al. 2012). The size of the effect that images had on comprehension for the ASD group was r = 0.144 and for the control group it was r = 0.06, where "a large effect is .5, a medium effect is .3, and a small effect is .1" (Fritz et al. 2012).

Given these results, we can safely conclude that the inclusion of images to accompany complex words in a text had almost no effect on comprehension for adult participants with high-functioning autism and for adult non-autistic participants, even when accounting for the small size of our sample.

***H.7**: **Within groups**, images included in the text do not have an effect on **comprehension** as measured through **subjective measures**.*

To test this hypothesis we compared the data obtained through the question *"Do you think the insertion of images in some of the texts helped you comprehend the text better?"*, with possible answers *"yes", "no"* and *"cannot say"*. The results for the ASD group showed that 72.22% of the participants felt that images did help them comprehend the text better, while only 11.11% felt they did not help and 16.66% were undecided (Figure 6.5).
For the control group, the prevailing opinion was that images did not help their comprehension (44.44%), with 33.33% of the people feeling that images

did help their comprehension and 22.22% were undecided.

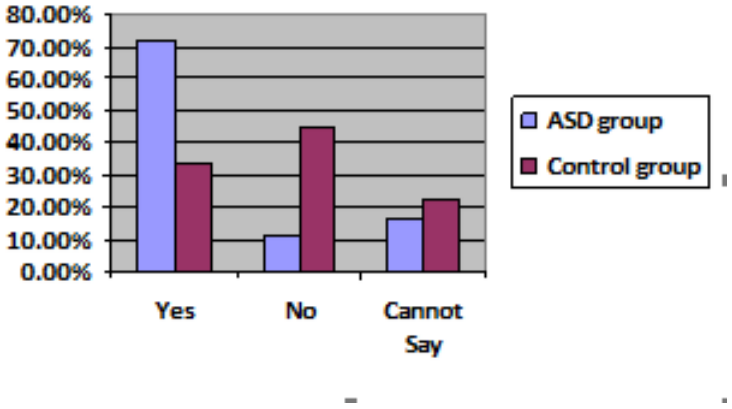

Figure 6.5: "Do you think the insertion of images in some of the texts helped you **comprehend** the text better?" Between-group comparison of the subjective effects of images on comprehension

6.5.5 Effects of Images on Memorisation and Recall

H.6: **Within groups**, images included in the text do not have an effect on **memorisation** as measured through **objective measures**.

To test whether images have an effect on memorisation and recall we compared the answers to the delayed multiple-choice questions (objective measures) for a condition, where a text is presented with images (Condition A) to the answers to the delayed multiple-choice questions (objective measures) for a condition, where the same text is presented without images (Condition B).

The results from a Mann-Whitney U test indicated that there was no

within-group effect for the ASD group (U = 1377, p = 0.545, two-tailed), nor for the control group of non-autistic participants (U = 1404, p = 0.542, two-tailed).

As in the previous subsection, we measured the size of the effect using Cohen's r (Cohen 1988), to make sure that the lack of an effect was not due to the small sample size. The results indicated that the effect was very small both for the ASD group (r = 0.1424) and for the control group (r = 0.1437), meaning that if the study were replicated with a larger sample the effect would remain small.

H.8: **Within groups**, images included in the text do not have an effect on **memorisation** as measured through **subjective measures**.

A within-group assessment of the data obtained through the question *"Do you think the insertion of images in some of the texts helped you memorise the text better?"*, with possible answers *"yes", "no"* and *"cannot say"* revealed that for the ASD group, participants felt that images aided the memorisation of text even more than they helped with comprehension (77.77% of the participants answered "yes"), with 5.55% of the participants giving a negative answer and 16.66% choosing the "cannot say" option (Figure 6.6).

As in the results for the comprehension question, the prevailing opinion among the control group was that images did not help their memorisation and recall (50%), with 38.88% feeling that images did help them memorise

the information and 11.11% undecided (Figure 6.6).

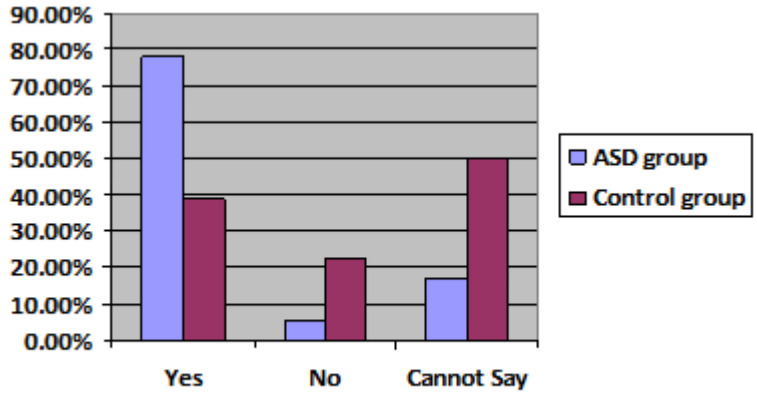

Figure 6.6: "Do you think the insertion of images in some of the texts helped you **memorise** the text better?" Between-group comparison of the subjective effects of images on memorisation

6.5.6 Between-group Differences in the Effects of Images on Comprehension and Memorisation

H.9: **Between groups**, images included in the text do not have an effect on **comprehension** as measured through **objective measures**.

A between-group comparison of the answers to the comprehension multiple-choice questions revealed a statistically significant difference between the level of comprehension of the ASD group and the control group, which was to be expected given the reading difficulties of people with autism, which are the subject of this article ($U = 5076$, $p = 0.028$, two-tailed).

H.10: **Between groups**, images included in the text do not have an effect on **memorisation** as measured through **objective measures**.

The difference between the level of text memorisation of the two groups was even more dramatic than their differences in comprehension, as confirmed by a Mann-Whitney U test ($U = 4482$, $p < 0.0001$, two-tailed). This is an interesting result to be explored further, as it suggests that some of the reading difficulties encountered by people with autism might be due to challenges related to retaining the information gained through reading and not solely due to challenges related to comprehending the text.

H.11: **Between groups**, images included in the text do not have an effect on **comprehension** as measured through **subjective measures**.

The answers for each group for this question were presented in Section 6.5.4; however, in this section we compare the two groups to highlight the different patterns of subjective perception exhibited by the groups (Figure 6.5). A Pearson Chi-square test revealed that the ASD participants perceived images as helpful cues to text comprehension significantly more often than the control-group participants ($\chi^2(2) = 8.023$, $p = 0.018$).

H.12: **Between groups**, images included in the text do not have an effect on **memorisation** as measured through **subjective measures**.

Figure 6.6 shows a dramatic difference between the subjective opinions of the participants from both groups on whether or not images aid memorisation and recall of information. 77.77% of the ASD group said that images did help them, while only 38.88% of the control group supported this statement. 50% of the control participants responded negatively. The difference between the subjective perceptions of the two groups on the effect of images on memorisation was a statistically significant one ($\chi^2(2) = 8.933$, $p = 0.011$).

6.5.7 Text-Presentation Preferences

H.4: There is no difference **between groups** regarding the text presentation preferences.

Although this hypothesis originally featured in the preliminary study involving easy-to-read documents, additional survey questions were subsequently added at the stage of the follow-up experiment in order to gain a broader understanding of the text-presentation preferences of users with autism. The question used in the preliminary study was: *"In your everyday life, do you prefer reading texts with: a) photographs, illustrating the main ideas b) symbols, illustrating the main ideas, c) plain texts without any images or d) "It makes no difference to me""*. There was a strong preference for the inclusion of images among the ASD group (58.81%), with 23.5% pre-

ferring texts with photographs and 35.3% preferring texts with symbols. The control group did not declare such a strong preference for images with 60% of the participants stating that it makes no difference to them and 30% in favour of the inclusion of images but undecided as to whether they preferred photographs (15%) or symbols (15%) (Figure 6.7).

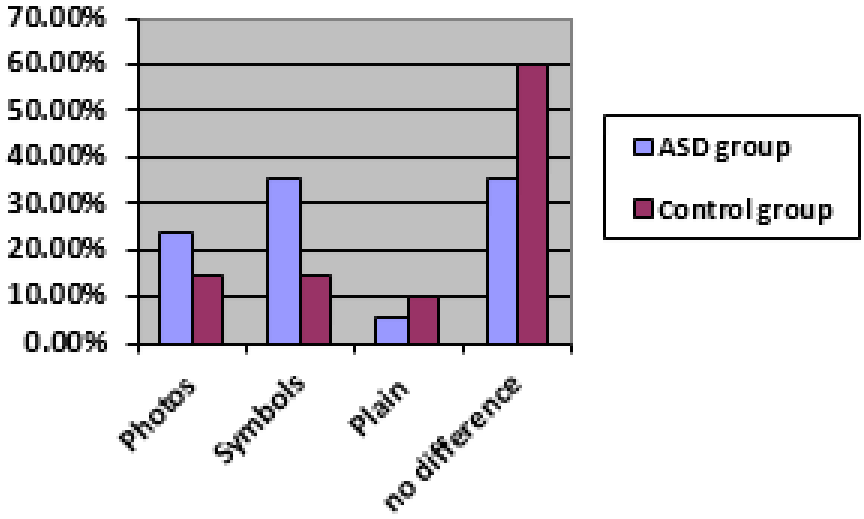

Figure 6.7: Preferences regarding the inclusion of images in text (initial study question)

In the follow-up study, where participants were presented both with texts containing images and texts not containing images, we included the following question: *"What did you find most useful: a) Reading text with images, b) Reading text without images or c) No preference"*. Within the ASD group, 72.2% of the participants answered that they preferred the texts with images as opposed to 16.66% who preferred those without images and 11.11% who had no preference. Compared to the control group the results once again

confirmed a clear preference for images among the participants with autism. Only 27.77% of the control participants said that they preferred reading texts with images and 11.11% that they prefer texts without images; however, the majority of the control participants (61.11%) responded that it makes no difference to them (Figure 6.8). The difference between the preferences of the two groups was statistically significant ($\chi^2(2) = 9.986$, $p = 0.007$).

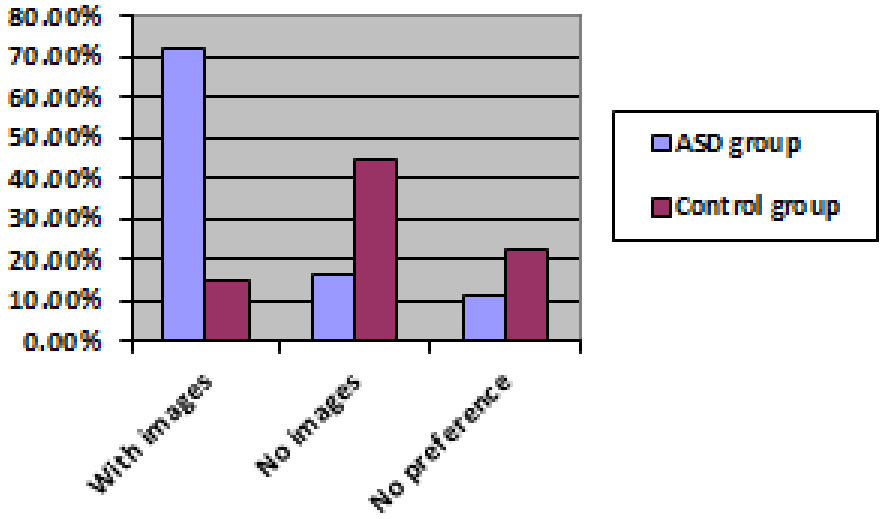

Figure 6.8: Preferences regarding the inclusion of images in text (follow-up study question)

Finally, we asked both groups in what position they would prefer to have an image illustrating complex words in texts. After showing them examples of how different versions looked, they had to choose between the following answers: *"a) On the right side of the sentence, b) Above the word, c) On the right side of the word, d) Other (please specify)"*. The majority of the

ASD participants chose having images right next to the word (77.33%), with 44% of the ASD participants choosing to have the image above the word and 33.33% of them choosing to have the image on the right-hand side of the word. 11.11% of the ASD group chose to have the images on the right-hand side of the sentence and another 11.11% selected the option "Other". One of them specified that he would prefer to have the image on the left-hand side of the sentence, so that it is the first thing the brain processes while the eyes move from left to right. The majority of the participants from the control group also chose to have the image inserted above the word (44.44%), with 38.88% of them choosing to have the image on the right-hand side of the word and 16.66% to have the image on the right-hand side of the sentence. None of the control group participants chose answer d) Other. All results from the user-preferences survey from both the initial and follow-up experiments consistently support the argument that users with autism strongly prefer reading texts containing images and that they perceive images as cues for improving their comprehension and memorisation of the text.

6.6 Discussion

The results presented in the previous section have several important implications concerning accessibility research, mainly with regards to attention, and comprehension and memorisation in readers with autism, as well as their user preferences in text presentation. This section presents a discussion of the main findings from these studies as well as a set of guidelines for text

accessibility for people with autism based on the experiments' results.

6.6.1 Methodological Challenges and Contributions

The rejection of **H1** confirmed that there were differences in the attention patterns between readers with and without autism with regards to the time they spent looking at images for each image and text pair. While this result is in line with previous research (Section 6.2.2), it poses the question of whether the readers with autism spent longer concentrating on the images because they used them as comprehension cues or because they were distracted by them. To test this further we conducted reading-comprehension experiments, where we investigated the effects images in text had on text comprehension and memorisation in readers with autism measured both objectively and subjectively. An interesting finding was the fact that images did not have any significant effects on the way readers with autism comprehended or memorised the meaning of a text; however, the behaviour of participants with autism demonstrated clearly that they perceived images as cues that helped them comprehend and memorise texts. This subjective perception was consistent when measured through four differently-worded questions featured in both Study 1 and Study 2; this perception also exhibited a pattern which was unique to the group with autism, since the control participants predominantly chose the answer "it makes no difference to me".

Interpretations of this result include the possibility that participants with autism were more prone to suggestion compared to the group without autism.

However, this interpretation would not explain the consistency of the data elicited by four differently-worded questions. Another interpretation could be the fact that images served as a natural way to segment the text into smaller portions, which helped readers with autism to assimilate the information more easily. Whichever way this result is interpreted, the conclusion of this study is that readers with autism have a strong preference for having images included in the texts they read.

As far as image type is concerned, there was no difference between the times participants focused on photographs and on symbols, suggesting that both types of images are equally suitable to be used in text documents for adult readers with autism. This finding is not in conflict with previous research, as it is the only study on images so far that has included adult autistic participants instead of children. It is possible that there would be a significant difference between the two sets if the participants were children, as it may be the case that symbolic understanding in autistic individuals reaches levels equal to those of neurotypicals later in their lives. In this sense, the results of this study with regard to the types of images preferred should not be generalised to children or to autistic individuals with learning difficulties.

The fact that texts written in plain English were perceived as ranging from very easy and easy to difficult and even very difficult by the participants with autism, while non-autistic participants rated them predominantly as very easy, is an indication that these texts were well understood by the autistic participants without being as trivial and under-stimulating to them as they

have appeared to be for the non-autistic ones. Perceived level of difficulty is not a direct measure of interest but one could hypothesise that texts which are too easy would bore the readers and thus reduce their motivation to read the document. The results suggest that even though our autistic participants were adults without a learning disability, this was not the case with them, and that Plain English is indeed a suitable level of difficulty for this population.

One factor which may have influenced these results is that the study did not use deception and all participants knew that it investigated reading in autism. The autistic participants might have suggested they were expected to have some sort of reading deficit and thus might have tried to apply a more fine-grained classification of the difficulty of the texts compared to the non-autistic ones. Nevertheless, differences in the interpretation of Likert items are a well-known flaw in all types of studies using this measure, while in the case of this study the results from the Likert scale were in agreement with the longer reading times of the autistic participants, which support the conclusion that they indeed did not find the texts as easy as the control participants did.

Another interesting result was the finding that while participants with autism scored lower on comprehension and memorisation when compared to participants without autism, there was a dramatic between-group difference in their answers to the memorisation questions. In other words, compared to the control participants, the readers with autism struggled much more with memorising the information than with comprehending it. This find-

ing suggests that some reading deficits among people with autism may be due to difficulties retaining the information and not solely to comprehension deficits. This finding implies that strategies aiming to aid comprehension should also place special emphasis on memorisation, as outlined in the accessibility guidelines at the end of this section.

6.6.2 Limitations

Limitations of both studies include the relatively small number of documents assessed and the small number of participants. The former was imposed by the difficulty experienced by autistic participants in concentrating for long periods of time. Owing to difficulties with concentration, the number of texts assessed and the number of questions featured for each of the documents were relatively small. The second limitation is typical of all areas of autism research, which makes the results from these studies difficult to generalise and is the reason for the many inconsistencies in autism study replications. The reason that the samples in ASD research tend to be small is the varying levels of ability of among people with autism and the number of comorbid disorders, (e.g. learning difficulties, dyslexia, depression, apraxia) which are so common among people with autism but which in many cases need to be excluded for the purposes of research.

One way in which we tried to address these limitations was through measuring the effect size for negative results in order to make sure that the lack of a significant result was not due to the limited size of our sample. In all such

cases, the effect of images was shown to be very small, which suggests that changes would be unlikely in the significance of this result if the experiment were repeated with a larger sample of participants.

The next section presents accessibility guidelines for people with autism based on the experimental results from the studies from this chapter.

6.6.3 Guidelines for Improving Text and Web Accessibility for People with Autism

A) Insertion of Images

1. Illustrate the main ideas in text paragraphs through the insertion of images relevant to the meaning of the paragraph. (Hypotheses 7, 8, 4)

2. Illustrate the complex words in the text through the insertion of images relevant to the meanings of words. (Hypothesis 4)

3. Even if a text does not contain complex words, it is still good practice to include images, as they have been shown to have a positive impact on how well autistic people perceive their comprehension and memorisation of the meaning of the text. (Hypotheses 7 and 8)

4. For texts containing many complex words, images should be accompanied by other comprehension aids such as dictionary look-up or the inclusion of definitions, where possible. This is needed because even though images have been shown to have a positive effect on the subjective perception of comprehension and recall among individuals with

autism, they have not been shown to improve them objectively. (Hypotheses 5 and 6)

B) Types of Images

5. Photographs and symbols are equally suitable, so data sets from both domains could be utilised. However, refrain from using symbols which are too abstract or whose understanding requires substantial prior knowledge in a certain area. (Hypothesis 2)

6. If a relevant image is unavailable or the idea of the text is too abstract to be depicted as an image, do not put anything. An irrelevant image has the potential to affect autistic readers' comprehension and reading speed.(Hypothesis 1)

7. Do not insert logos, advertisements or any other visual information, which is not directly relevant to the meaning of the text. (Hypothesis 1)

C) Positioning of Images

8. If an image has been used to illustrate a complex word, position the image as close as possible to the word, preferably above the word or on

the right-hand side of the word. (Hypothesis 4)

9. If an image has been used to illustrate the meaning of a larger portion of text, insert the image as close as possible to the sentence or groups of sentences it refers to. (Hypothesis 4)

10. Images should be positioned in a way that aids the natural segmentation of the text, as opposed to segmenting the text again and, in so doing, interfering with meaning and cohesion. (Hypothesis 4)

D) Supporting Comprehension

11. Use texts written in Plain English. See Plain English guidelines for a more detailed information on how to write for people with cognitive disabilities (Tronbacke 1997). A general rule of thumb is that the text should have a score higher than 65 according to the Flesch-Reading Ease formula (Flesch 1948). (Hypothesis 3)

12. Allow re-reading of the text as some readers might need to read it several times in order to comprehend and memorise it fully. (Hypothesis 3, part two)

13. Reinforce prior knowledge on the subject by asking a few questions about the topic before the text has been read.

14. Reinforce comprehension by asking inferential questions after the text has been read.

E) Supporting Memorisation

Supporting memorisation through various means is necessary due to the fact that the between-group difference in memorisation was more dramatic than the between-group difference in comprehension (Hypothesis 10). At the same time, it was shown that image insertion does not affect momerisation objectively, which is why we propose the following practices [7]:

15. Reinforce memorisation of important information in the text by presenting a summary of it after the text has been read.

16. In the case of instructions, reinforce the information by displaying the relevant chunks of previously read text while the user is taking the required action. For example, if a text explains to a user how to create an account, the text should first be displayed at the beginning of the process and then each step of the registration process should be accompanied by the relevant description outlined in the beginning.

[7]It is important to note that the effectiveness of those has not been empirically tested

F) Reading Speed

17. Allow readers to skip through pages at their own pace, as their reading time may be longer compared to the general population. (Hypothesis 3)

18. Reinforcing important information (see Hypothesis 10) by presenting a summary of it at the end of the text will increase text length and will also have an impact on the overall reading time. This needs to be taken into account in situations such as online gaming or videos, where the trade-off between quality of comprehension and speed of comprehension is important.

19. In the case of videos, allow longer for the users to read the text or captions and to process the visual information. (Hypothesis 3)

6.7 Summary

This section presented studies into the effects images in text have on the attention, comprehension, memorisation and user preferences of readers with autism. The main findings showed that participants with autism tend to focus on images in text for longer periods of time compared to the participants without autism; however, images did not have a significant effect (either positive or negative) on their comprehension or memorisation of the information in the text. Unlike the control group, the participants with autism had a

very strong preference towards having images inserted in the texts they read and subjectively perceived these texts as easier to comprehend and memorise. Additional findings included: Photographs and symbols are equally suited for inclusion in documents for adults with autism, and there were more dramatic between-group differences in the memorisation of the information than in its comprehension. The latter suggests that accessible texts for people with autism should place a particular emphasis on reinforcing memorisation of information.

Finally, this chapter presented text-accessibility guidelines for people with autism, based on the results from the experiments presented above.

The next chapter presents a study into the way people with autism search for information within web pages.

Chapter 7

Web Searching in Users with Autism: Do Barriers to Finding Relevant Information Exist?

7.1 Chapter Overview

In previous chapters, we discussed text readability and text presentation, particularly with regards to the inclusion of images in text, as a way to improve the reading comprehension of people with autism. In the 21st century, where we look for information has shifted from the printed page (newspapers, textbooks) to the Internet. This shift requires new skills for information searching, such as formulating a query and identifying the relevant links, scanning web pages, ignoring adverts while focusing on reading the relevant information, coping with a large amount of visual stimulation (e.g. images, adverts, videos), etc. Research from the field of Human-Computer Interaction (HCI) aims to cast light on the issues web users experience when interacting with technology; however, until now, the way users with autism interact with the web has not been empirically investigated.

In this experiment we build upon our findings on the effects of images on attention in readers with autism, presented in Chapter 6, by investigating the

effects of visual complexity on the success of web users with autism in finding relevant information on web pages. We test whether web users with autism perform differently when presented with pages of low, medium and high visual complexity. We compare the performance of two groups of participants (those with and those without autism) when searching for information on web pages while also comparing gaze data collected during their search. The aim of the experiment presented in this chapter is to establish whether adults with autism face any barriers when searching for information within web pages and if so, what these barriers are. These experiments address RQ5:

RQ5: Do web users with autism encounter barriers to finding information on web pages?

The empirical evidence for barriers individuals with autism encounter when searching for information within web pages is considered to be the fifth original contribution of this thesis.

Some of the experiments in this chapter (the ones referring to Hypotheses 1 and 4) have been presented in Eraslan et al. (n.d.) (under review).

7.2 Motivation

7.2.1 Autism and Web Accessibility

Autism and cognitive disabilities in general have been shown to have an impact on the way users interact with the web. This impact is due to the following issues: "difficulty in using the web due to limited reading comprehension, complexity, slower learning, limited fine motor control (...) and lowered information overload threshold" (Friedman & Bryen 2007). Other related challenges include "difficulty in recognizing the most appropriate choice when faced with a large number of options and distinguishing foreground images and text from background material" (Slatin & Rush 2003). These findings are supported by research from the field of psychology, which describes certain aspects of the autism profile with a bias in favour of processing local sensory information, with less account for global, contextual and semantic information (Weak Central Coherence Theory (WCCT)) (Happé & Frith 2006). According to WCCT, in the context of searching for information within web pages, people with autism would be expected to focus more on potentially irrelevant details, which prevent them from perceiving the bigger picture. WCCT is in line with the stimulus over-selectivity phenomenon in autism (Lovaas & Schreibman 1971), where part of the sensory information is neglected, causing "tunnel vision", a focus on detail to the detriment of the bigger picture (Ploog 2010).

As discussed in Chapter 1, the most widely used web accessibility guide-

lines which aim to meet the needs of all disabled user groups is WCAG 2.0 by W3C/WAI group (Caldwell et al. 2008). However, issues related to cognitive disabilities have been assigned lower priorities within these guidelines (Britto & Pizzolato 2016) and have been least discussed in WCAG and the literature (Harper & Yesilada 2008). One reason for the lack of clearer instruction regarding this user group is the virtually non-existent scientific investigation of the way they interact with the web, with the exception of a pilot study by Deering (2013) involving four participants with autism. Britto & Pizzolato (2016) compare existing guidelines relevant to web accessibility for people with autism by grouping them into the following categories: engagement, affordance, customisation, redundant representation, multimedia, feedback, system status, navigability and interaction with a touch screen. Unfortunately, these guidelines have not been empirically tested and supported with scientific studies.

7.2.2 Visual Complexity of Web Pages

The complexity of a website's presentation depends on the way its pages are designed, what elements are used and how information is grouped and positioned (Michailidou 2006). Based on a qualitative analysis of web-page attributes, Ivory et al. (2001) show that page-composition metrics, such as word count, number of images, tables and links, could distinguish between pages with good accessibility and pages with bad accessibility. Other studies identify an implicit link between visual complexity and cognitive complex-

ity, approaching visual complexity as an implicit measure of cognitive load (Michailidou et al. 2008).

A common approach to evaluating website accessibility and usability from the perspective of users is eye tracking, which enables investigation of the scan paths users follow to find given information and the order in which they fixate various elements of the web page. By using gaze data, Pan et al. (2004) show that website-viewing behaviour is determined by gender, the order of web pages being viewed and a possible relationship between scanpath variability among individuals and the structural/visual complexity of the web page (Pan et al. 2004, Michailidou 2006).

In this experiment we extend our findings on the effects of images on reading in autism presented in Chapter 6 to an investigation of the effects of visual complexity on the success of web users with autism in finding relevant information on web pages. We test whether web users with autism perform differently when presented with pages of low, medium and high visual complexity.

7.2.3 Study Aims

The aim of the study presented in this chapter is to investigate the way adults with high-functioning autism search for information within web pages. The study design, materials and procedure replicate an existing study by Eraslan & Yesilada (2015), which collected data on web-searching behaviour in neurotypical web users. Although we follow the same design and procedure as

Eraslan & Yesilada (2015), the purposes of the two studies are different. The objective of the study by Eraslan & Yesilada (2015) was to use the collected data in order develop the eMINE scan-path algorithm which concatenates scan paths from multiple users. The purpose of our study, however, is to compare the search performance of two groups of participants (those with and those without autism) by using results from comprehension questions and gaze data. We measure participants' success in locating the required information on the web page and we analyse the gaze data to identify the visual elements that may have caused differences in attention and performance. In this chapter we focus on investigating web accessibility purely from the perspective of visual elements (such as headers, footers and hyper links) and their organisation within web pages.

This rest of this chapter is organised as follows. Sections 7.3 and 7.4 present the design and hypotheses of the research study. Then, we discuss the experimental set up including participants, materials, apparatus and procedure. Section 7.6 presents the results from the study, which are then discussed in Section 7.7.

7.3 Design

This study employed a between-group comparison design comparing the performance of 18 participants diagnosed with high-functioning autism and 18 control neurotypical participants on a web-search task. While performing

the task, their eye movements were recorded by an eye tracker.

Each participant was shown six web pages with different levels of visual complexity and asked to answer two questions per web page (12 questions in total) about finding relevant information on the page. The web pages, as well as the questions pertaining to each page are presented in Section 7.5.1. The time limit for answering the two questions for each web page was 30 seconds.

Once the task was completed, the participants rated their familiarity with the websites where the pages were taken from (Apple[1], Babylon[2], AVG[3], Yahoo![4], GoDaddy[5], and BBC[6]).

After that, they were asked to fill in a short survey containing the following questions: *"How often do you use the web?"; "How easy or difficult is it for you to find the information you need when you search the web?"* and *"When you search for something in Google (or other search engines), how easy or difficult is it for you to know which links to open in order to find the information you need?"*. The latter question was based on the findings from previous research stating that people with cognitive disabilities may have difficulty: "recognising the most appropriate choice when faced with a large number of options" (Slatin & Rush 2003). The answers to these questions

[1]Apple. Available at: http://www.apple.com/uk/

[2]Babylon Translation. Available at: http://translation.babylon-software.com/

[3]AVG. Available at: http://www.avg.com/gb-en/homepage

[4]Yahoo! Available at: https://uk.yahoo.com/

[5]GoDaddy. Available at: https://uk.godaddy.com/

[6]BBC news. Available at: http://www.bbc.co.uk/

were measured using five-point Likert scales.

The next section presents the hypotheses tested in this study.

7.4 Study Hypotheses

The formal hypotheses tested in this experiment are as follows:

H1: Between-groups, there is no difference in the success of correctly locating information on the web pages in response to 12 web-search questions.

This hypothesis tests whether the participants with ASD had actual difficulties searching for information on web pages as compared to the control group of neurotypical participants.

H2: Between-groups, there is no difference in the average number of fixations, revisits and overall dwell time in a web-page search task.

Since the number of fixations and their duration is an indication of higher cognitive load, this hypothesis indirectly tests whether there are differences in the cognitive effort made by the two groups of participants in completing the task.

H3: Within groups, the level of visual complexity of the web pages does not have an effect on the success of correctly locating information on the web pages.

This hypothesis investigates the effects of visual complexity on the success of finding information within web pages, potentially leading an improvement in web accessibility for users with autism through a reduction in visual com-

plexity.

H4: Between groups, there is no difference between the perceived level of difficulty of searching the web.

This hypothesis refers to the data obtained from the survey questions and aims to account for the subjective experiences of users with autism when searching the web.

The next section presents the experimental methodology.

7.5 Method

This section presents the materials, participants, apparatus and procedures used in this study.

7.5.1 Materials

The materials used in this study were the screen shots of six web pages that were initially selected for and used in Eraslan & Yesilada (2015). These web pages had been selected from a list of top websites by traffic at ALEXA.com[7]. The six web pages had varying visual complexity (low, medium, high), as measured by the ViCRAM tool (Michailidou 2010): Apple (Low), Babylon (Low), AVG (Medium), Yahoo (Medium), Godaddy (High) and BBC (High). Figures 7.1, 7.2, 7.3, 7.4, 7.5, and 7.6 show the web pages used as stimuli in the experiment.

[7] `http://www.alexa.com`

Figure 7.1: Screenshot of the Apple web page (low visual complexity)

Figure 7.2: Screenshot of the Babylon web page (low visual complexity)

Figure 7.3: Screenshot of the AVG web page (medium visual complexity)

Figure 7.4: Screenshot of the Yahoo! web page (medium visual complexity)

Figure 7.5: Screenshot of the GoDaddy web page (high visual complexity)

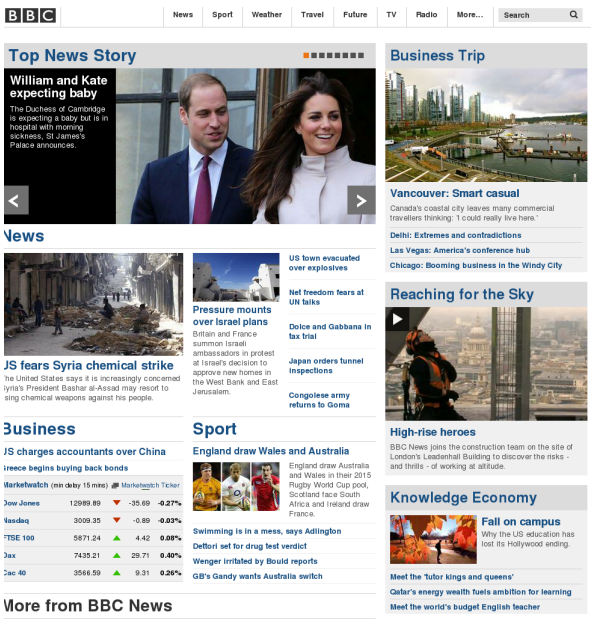

Figure 7.6: Screenshot of the BBC web page (high visual complexity)

Table 7.1: List of search tasks for the six web pages

Apple
(a) Can you locate the link that allows watching the TV ads relating to iPad mini?
(b) Can you locate a link labelled iPad on the main menu?
Babylon
(a) Can you locate the link that you can download the free version of Babylon?
(b) Can you find and read the names of other products of Babylon?
AVG
(a) Can you locate the link which you can download the free trial of AVG Internet Security 2013?
(b) Can you locate the link which allows you to download AVG Antivirus Free 2013?
Yahoo
(a) Can you read the titles of the main headlines which have smaller images?
(b) Can you read the first item under the News title?
Godaddy
(a) Can you find a telephone number for technical support and read it?
(b) Can you locate the text box where you can search for a new domain?
BBC
(a) Can you read the first item of Sport News?
(b) Can you locate the table that shows market data under the Business title?

Table 7.1 presents the tasks the participants had to solve for each web page.

The areas of interest were defined by Eraslan & Yesilada (2015) based on the Vision-Based Page-Segmentation (VIPS) algorithm, which segments web pages by using their source code and visual representations based on different granularity levels (Akpınar & Yeşilada 2013*a*, Akpınar & Yesilada

2013*b*). For the web pages used in this experiment, Eraslan & Yesilada (2015) selected the fifth granularity level, owing to the fact that "it was found as the most successful level with approximately 74% user satisfaction" (Eraslan & Yesilada 2015). Figures 7.7 and 7.8 show examples of the areas of interest for the Apple and BBC web pages, respectively.

Figure 7.7: Areas of interest on the Apple web page

Figure 7.9 shows the scan paths of one participant with ASD (purple) and one control group participant (green) for the Yahoo! web page, while answering the question: "Can you read the titles of the main headlines which have smaller images?". The next subsection will present details regarding the participants in this study.

Figure 7.8: Areas of interest on the BBC web page

7.5.2 Participants

The participants in the study were 18 adult volunteers diagnosed with high-functioning autism or Asperger's syndrome (12 male and six female) and 18 non-autistic control participants (ten male and eight female). The inclusion and exclusion criteria for the two groups were the same as those outlined in Chapter 3. All participants had normal or corrected vision. The mean age for the ASD group was m = 37.22 with standard deviation SD = 10.3 and for the control group, the mean age was m = 34.18, with standard deviation SD = 8.05. The number of years spent in education for the ASD group was m = 16, SD = 3.33 and for the control group was m = 18.35, SD = 2.47. Three

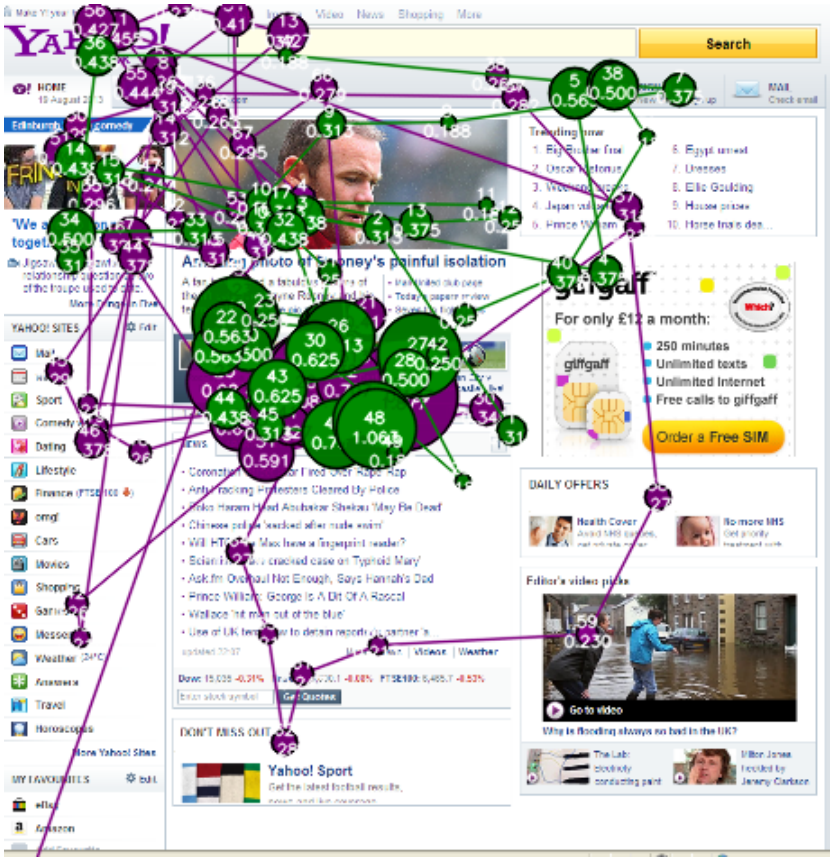

Figure 7.9: Scan paths of one participant with ASD (purple) and one control group participant (green) for the Yahoo! web page.

of the ASD-group participants were subsequently excluded from the analysis of the eye-tracking data due to their inability to calibrate the device.

All participants were regular web users who reported the following familiarity levels with the websites from which the web pages were taken. For the AVG website, 82.35% of the ASD participants reported that they had never visited it and 17.65% reported they visited it less than once a month; for the Apple website 52.94% of the ASD participants reported never having visited it and 47.06% having visited it less than once a month; none of the ASD participants had ever visited the Babylon website; the BBC website obtained more diverse responses, with 11.76% visiting it less than once a month, 23.53% - monthly, 23.53% - weekly, and 41.18% - daily. Among ASD participants, 94.12% had never visited the GoDaddy website and 5.88% of them would visit it less than once a month; finally, for Yahoo!, 23.53% of the ASD participants had never visited it, 29.41% - visited less than once a month, 17.65% - visited monthly, 11.76% - visited weekly, and 17.65% - visited daily.

The percentages for the control group for the AVG website were: 62.5% - never, 31.25% - less than once a month, and 6.25% - monthly. The Apple and BBC websites were the most visited ones with only 12.5% of the control participants never having visited them, 25% - visiting less than once a month, 37.5% - visiting monthly, and 25% - visiting weekly. 81.25% of the control participants had never visited the Babylon website and 18.75% of them visited it less than once a month. The figures for GoDaddy were: 75% - never,

18.75% - less than once a month, and 6.25% - monthly. Finally, 43.75% of the control participants had never visited Yahoo!, 43.75% - visited less than once a month, 6.25% - visited monthly and 6.25% visited weekly.

7.5.3 Apparatus

The device used for recording the gaze of the participants while they performed the tasks was a Gazepoint GP3 video-based eye tracker (Gazepoint 2015) (60Hz sampling rate and accuracy of 0.5 - 1 degree of a visual angle). The screenshots of the web pages were displayed on a 19" LCD monitor. The eye tracker was calibrated individually for each participant using a 9-point calibration procedure. The distance between each participant and the eye tracker was controlled using a sensor integrated within the Gazepoint software, and was roughly 65 cm.

7.5.4 Procedure

The experiment was performed in a quiet room with only the researcher present. All participants were familiarised with the purpose and procedure of the experiment and signed a consent form. Demographic data (age, gender, diagnoses) were collected and a nine-point calibration of the eye tracker was carried out. After that, participants were presented with the six web pages in a randomised order and given 30 seconds to answer the two questions for each page. Participants were not required to use mouse or keyboard. Once

the task was completed for all six web pages, data on website familiarity were collected, the survey questions were answered and participants were debriefed.

7.6 Results

H1: Between-groups, there is no difference between the success of correctly locating information on the web pages in response to 12 web search questions.

This hypothesis was tested by comparing the answers to the web-search questions, where each correct answer was given a score of 1 and each incorrect answer- a score of 0. A Chi-square test for independence (Pearson's Chi-square test) indicated that there was a statistically significant difference between participants with and without autism in their success of locating the required information on the web pages ($\chi^2(1) = 4.780$, $p = 0.029$). This result suggests that adults with autism are less successful in locating the correct information on a web page under limited time constraints.

H2: Between-groups, there is no difference in the average number of fixations, revisits and overall dwell time in a web page search task.

First, we applied the Shapiro-Wilk test for normality of distributions, which confirmed that the data for both the ASD and the control groups were non-normally distributed ($p = 0.00$) for all AOIs in a search task. We then applied the Mann-Whitney U test for non-parametric data, which showed that

there was no statistically significant difference between the average number of fixations for the two groups (U = 5683.5, $p = 0.224$). We did the same comparison for the Average Time Viewed (ATV) and Average Revisits (AR) measures, which were both shown to be non-normally distributed according to the Shapiro-Wilk test. The Mann-Whitney U test once again confirmed lack of statistically significant difference between the two groups for both the ATV (U = 5950, $p = 0.506$) and the AR (U = 5563, $p = 0.140$) measures.

H3: Within groups, the level of visual complexity of the web pages does not have an effect on the success of correctly locating information on the web pages.

To test this hypothesis, we used the Wilcoxon Signed Rank test with Bonferroni correction of the significance level adjusted to $\alpha = 0.017$.

For the ASD group, the results indicated that there were statistically significant differences between the success of correctly allocating information within web pages with *low* versus *medium* levels of visual complexity (Z = -2.887; $p = .004$, two-tailed), as well as with *medium* versus *high* levels of complexity (Z = -3.357; $p = .001$, two-tailed). What was surprising however, was the fact that there was no statistically significant difference between web pages with *high* versus *low* visual complexity (Z = -1.732; $p = .083$, two-tailed), indicating that the web pages with *medium* complexity were actually the most difficult for the participants with autism.

A similar result was observed for the control group, where the only sig-

nificant difference was between web pages with *medium* versus *high* visual complexity (Z = -2.449; $p = .014$, two-tailed) owing to the fact that the pages with *medium* complexity were more difficult. There were no significant differences between *low* versus *medium* (Z = -2.000; $p = .046$, two-tailed) or *high* versus *low* (Z = -1.414; $p = .157$, two-tailed).

The results did not refute **H3**, thereby indicating that visual complexity, as measured in this experiment, does not have an effect on success in finding relevant information.

H4: Between groups, there is no difference between the perceived level of difficulty of searching the web.

The first survey question asked was *"How easy or difficult is it for you to find the information you need when you search the web?"*. 77.77% of the control group affirmed it was "very easy" for them to find the necessary information compared to 66.66% of the ASD group. In both groups, 16.66% of the users selected the option "easy", while the option "medium" was selected by 11.11% of the control group and 5.55% of the ASD group. While all answers of the control group spanned from "very easy" to "medium", 5.55% of the ASD group reported that they find it "very difficult" to find the information they need when they search the web (see Figure 7.10). However, Pearson's Chi-Square test revealed that there was no statistically significant association between the type of the participants (ASD or control) and their perceived difficulty searching the web ($\chi(3) = 1.487$, $p = 0.685$).

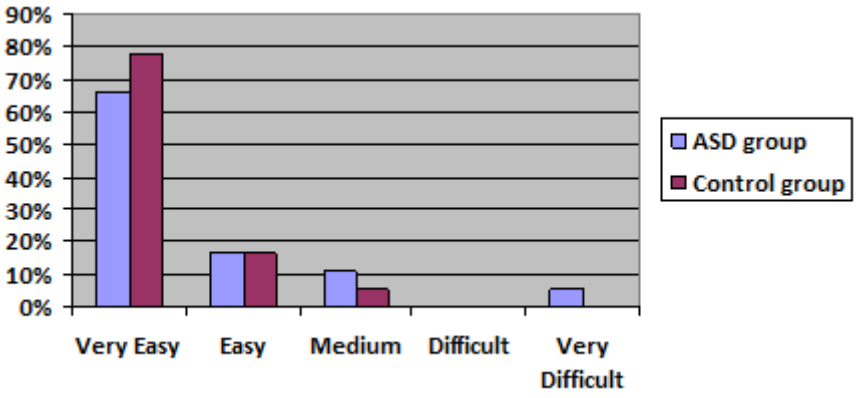

Figure 7.10: "How easy or difficult is it for you to find the information you need when you search the web?"

The second survey question was: *"When you search for something in the web, how easy or difficult is it for you to know which links to open to find the information you need?"*. The answers of the participants are provided in Figure 7.11.

As with the previous survey question, 77.77% of the users in the control group and 66.66% of those in the ASD group reported that it was "very easy" for them to select a relevant link. 16.66% from both groups selected the option "easy", and 5.55% selected "medium". However, 11.11% of the ASD group reported that this task was "very difficult" for them, compared to 0% of the control group. This is an indication that certain autistic users tend to find it challenging to filter information when it comes to web search. Nevertheless, the association between autistic traits and difficulties selecting relevant links was not statistically significant according to Pearson's Chi-

square test ($\chi(3) = 2.154$, $p = 0.541$)

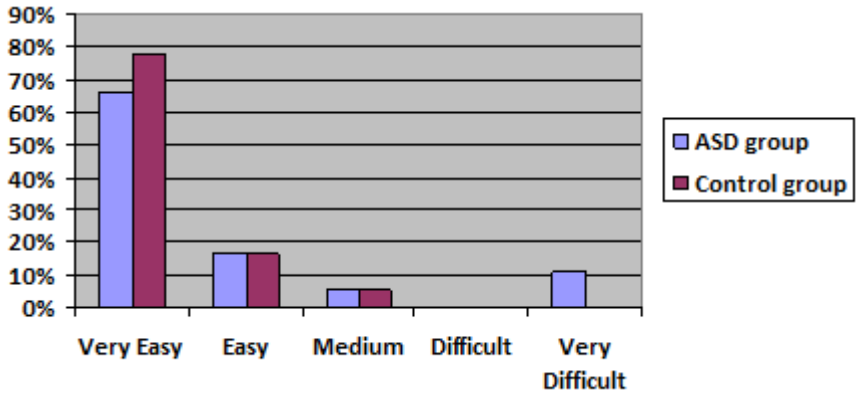

Figure 7.11: "When you search for something in the web, how easy or difficult is it for you to know which links to open to find the information you need?"

7.7 Discussion

This section discusses the implications and impact of the results presented in Section 7.6.

7.7.1 Methodological Challenges and Contributions

The finding that web users with high-functioning autism do experience barriers to searching for information within web pages is, to the best of our knowledge, the first empirical evidence for such difficulties among people on the autism spectrum. Since differences in performance were registered for participants with the mildest form of autism, where intellectual abilities are

intact, it could be hypothesised that the performance of web users with more severe forms of autism would be even more severely impaired. Accessibility barriers could be even greater if the target information on the web page is contained in written text (as opposed to in images or hyperlinks), where reading comprehension deficits add another layer of difficulty. Further research is needed to investigate web accessibility for people with autism from the perspective of a reading task, given that web text differs from other forms of written text, owing to the presence of hyperlinks, different navigation structures, different layouts and different organisation of information.

While the data analysis presented evidence for barriers to searching for information within web pages for users with autism, the results did not provide specific insight into where these difficulties lay. There were no statistically significant differences found in the number of fixations, revisits and overall dwell time among the two groups; hence, there is no evidence for differences in the cognitive effort made by the two groups of participants when completing the tasks. Different levels of visual complexity of the web pages also did not affect their performance, with the exception of the web pages with medium complexity, which were actually judged as most complex by both groups of participants. Thus, the presented results are inconclusive as to whether the lack of significant differences (between pages with low and high visual complexity) was due to lack of effect of visual complexity on accessibility or due to measuring inaccurately the complexity of the stimuli used in this study. The significant differences observed between pages with medium

complexity compared to those with low complexity indicate that the latter is more likely to be the case. Hence, it is recommended that the study be replicated with a higher number of stimuli, the complexity of which should ideally be pre-evaluated by humans instead of by algorithms, as it was in the current study.

Finally, the data from the survey questions revealed that even though the majority of both autistic and non-autistic participants felt that it was relatively easy for them to find the information they needed when searching the web and to know which links to open, there was a portion of autistic participants, who reported levels of perceived difficulty below "easy" (overall 16.66% of the high-functioning ASD participants selected "medium" or "very difficult" compared to 0% of the control group).

7.7.2 Limitations

The lack of significant differences in the gaze measures between the two groups could be attributed to the small number of participants and of stimuli; hence, it would be beneficial to replicate the studies with a larger group of users and a larger stimuli sample. Furthermore, the results of the effects of visual complexity on the success of locating relevant information were not conclusive, given that the pages with medium complexity were actually the most difficult ones. A possible reason for this result could be the inaccurate assigning of complexity levels by the ViCRAM algorithm (Michailidou 2010); hence, a replication with a set of web pages pre-evaluated by humans is

recommended. Another limitation is the time limit of 30 seconds for the completion of the two tasks per page. This limit might have put more pressure on the participants with autism compared to the control-group participants and could thus have presented a potential bias in the results. However, limiting the time was necessary in order to obtain a measurement sensitive enough to the difficulties of the two groups; allowing unlimited time for task completion would have resulted in 100% correct answers from both groups, without accounting for the difficulties of the ASD participants.

Future work is needed to investigate the web-searching strategies of people with more severe forms of autism, as well as their coping with web pages which contain a lot of textual information.

7.8 Summary

This chapter presented experiments investigating the way web users with autism process web pages. It was shown that the participants with autism found it significantly more difficult to find relevant information on the web pages compared to the control-group participants, as evidenced by their lack of success in solving information-location tasks. In spite of providing empirical evidence for such difficulties, the data analysis did not reveal differences between the fixations, revisits and average viewing time of the two groups and did not provide conclusive evidence about the role of visual complexity in information searching among web users with autism (visual complexity did

not have effect on task performance). The survey questions revealed that the participants with autism found it slightly more difficult to search the web.

The next section presents the main conclusions of this thesis, their impact on accessibility research and avenues for future work towards developing more accessible texts and web pages for people with autism.

Chapter 8

Conclusions and Future Work

The main goal of this thesis was to investigate new ways to evaluate and improve text and web accessibility for adults with autism. We worked towards achieving this goal by:

- Exploring ways to evaluate automatically the accessibility of text content for readers with autism
- Investigating the effects of images on text comprehension and memorisation in readers with autism, inspired by the idea of relying on their strengths (e.g. strong visual thinking) to overcome their reading-comprehension difficulties
- And investigating the accessibility of web pages for users with autism

This chapter summarises the main results presented in this thesis, discusses their impact on accessibility research and highlights possible avenues towards more accessible future text documents and web content.

8.1 Text Readability

The main issue with the development of readability classifiers for people with autism was the lack of user-evaluated resources that could serve either as a gold standard for accessible writing for autistic people or as an extrinsic evaluation set for the classifiers.

In Chapter 3 we addressed this problem by developing the ASD corpus. The ASD corpus consists of 27 documents, whose readability was evaluated by 27 different adults with autism. The ASD corpus is the first of its kind to contain texts with comprehension scores obtained from adult participants with and without autism as well as gaze data obtained from the two participant groups. Both the comprehension scores and the gaze data contained in the corpus allow not only for the evaluation of text and sentence complexity (the initial purpose for which it was complied) but also for comparison between the two groups. The development of the ASD corpus addressed research question one:

RQ1: How can we obtain a collection of texts with known levels of difficulty for readers with autism?

The resulting ASD corpus was used as an unseen user-evaluated test set for the document-level readability classifier, presented in Chapter 4.

We modelled text complexity based on a set of 43 features, which were either matched to the reading difficulties of readers with autism (based on the

literature review from Chapter 2) or were shown to have a high discriminatory power in other readability studies described in the same chapter. We carried out a feature selection, which revealed that the most discriminatory features for the task of document-level readability classification were *Polysemous type ratio, Words per sentence, Fog index, Average sentence length, Age of acquisition of words, Second pronoun incidence, Imagability* and *Flesch-Kincaid Grade Level*.

We then followed a supervised machine-learning approach for the training of the classifier and evaluated both its internal validity by means of cross-validation but also its generalisability over user-evaluated unseen data (the ASD corpus). When evaluated on the unseen data, the classifier achieved 77% accuracy for distinguishing between *easy, medium* and *difficult* texts for readers with autism and 90% for ten-fold cross-validation. This was a significant improvement on the Flesch-Kincaid baseline, which achieved 52% accuracy. This classifier is the only attempt, as of yet, to measure text complexity automatically for people with autism, that has been evaluated on user data. The development of the document-level readability classifier addressed research question two:

RQ2: Is it possible to develop an automatic *document*-level readability classifier for people with autism, which generalises over unseen user-evaluated data better than existing readability metrics?

We further explored the problem of readability assessment for readers with autism at sentence level (Chapter 5). The motivation behind this was the fact that text-simplification systems need to have a way of selecting which sentences are complex and need simplification, thus leaving the rest of the sentences intact. Unlike previous studies on sentence-level readability assessment where the gold standard were sentences simplified by experts, we developed a gold standard of *easy* and *difficult* sentences based on the gaze data obtained from the participants with autism. In addition, we featured a set of 100 sentences with a control length, the complexity of which was evaluated through multiple-choice questions.

We modelled sentence complexity on a set of 47 features, grouped into categories of "shallow descriptors", "features of cohesion", "cognitively-motivated features" and "incidence counts". After the feature-selection process, only the 12 features with the highest discriminative power were retained for the final model. These were: *Word count, Word length in syllables, Word length in letters, Word length in letters (SD), Intentional verbs incidence, LSA verb overlap, Pronoun incidence, CELEX word frequency (mean), Concreteness (mean), Imagability (mean), Polysemy (mean)* and *Hypernymy for nouns (mean).*

The model was evaluated intrinsically using ten-fold cross-validation and achieved 82% accuracy; this was only a slight improvement on the baseline of sentence length, which achieved 78%. The development of the sentence-level classifier addressed research question three:

RQ3: Is it possible to develop an automatic *sentence*-level readability classifier for people with autism, that performs better than existing readability metrics?

The next subsection presents the impact the ASD corpus and the two readability classifiers may have on future research and on readers with autism.

8.1.1 Impact

The findings from these chapters and the readability classifiers described in them (Original Contributions 2 and 3) have the potential to change the way accessible texts for people with autism are developed by humans (e.g. easy-to-read documents or educational texts) and on text simplification systems. Both classifiers are in the process of being implemented in a tool called AUTOR, which makes more efficient the development and evaluation of accessible materials. AUTOR also has the potential to change the way students with autism are assessed in the education system if exam materials are evaluated and improved using the tool. The sentence-level readability classifier is particularly relevant to text-simplification systems because it helps to identify sentences in the text that need simplification and those that may be left as they are, thus reducing the workload of the system and the amount of human post-processing required to fix the grammaticality of the simplified output.

The developed ASD corpus (Original Contribution 1) has the potential

to change the types of research undertaken in the future. This influence could extend to the fields of readability assessment and text simplification (by using the data as an evaluation set) as well as to the field of clinical linguistics, by comparing gaze data and comprehension scores obtained from adults with and without autism.

Last but not least, these results have important implications for multidisciplinary research. The most informative features retained for both the document-level and sentence-level readability classifiers were very well-matched to the main reading difficulties of people with autism described in clinical linguistics and presented in Chapter 2. For example, the selected features for the document-level classifier were *Polysemous type ratio, Words per sentence, Fog index, Average sentence length, Age of acquisition of words, Second pronoun incidence, Imagability* and *Flesch-Kincaid Grade Level*. These features refer to reading difficulties quite typical for readers with autism such as difficulties in resolving ambiguity in meaning (polysemy), in processing long sentences, in resolving pronouns, as well as in abstract thinking and coping with unfamiliar words. It is important to note that many of these features (e.g., imagability or second pronoun incidence) have not been shown to be as relevant to other user groups (e.g. foreign language learners or children) as described in Chapter 2. The fact that the readability classifiers were actually able to capture text properties relevant to the autism profile was also supported by the selected features for the sentence-level classifier such as *Intentional verbs incidence, Pronoun incidence, Concreteness*

(mean), Imagability (mean) and *Polysemy (mean)*, among others.

It was heretofore believed that NLP approaches to readability and text simplification could be informed by psycholinguistic research, but we have also shown that this relationship could work both ways. Despite not using matched-group design, which the majority of studies in clinical linguistics do, we show that NLP techniques could also be well-suited to inform future research avenues in clinical linguistics.

8.2 Images in Text

Inspired by one of the main strengths of the ASD cognitive profile, namely strong visual thinking, we explored whether this strength could be used to compensate for the reading-comprehension difficulties people with autism experience. The main findings regarding the use of images in text and the preferences of adults with autism towards text presentation were discussed in Chapter 6 and address research question four:

RQ4: Do images inserted into texts have an effect on participants' attention, comprehension and memorisation of a text, measured both objectively and subjectively?

The experimental results showed that there were differences between the attention patterns of readers with and without autism with regards to the time they spent looking at images for each image and text pair. This finding

raised the question of whether the readers with autism spent longer concentrating on the images because they used them as comprehension cues or because they were distracted by them. To explore this question further, we carried out a follow-up experiment, which showed that images did not have any significant effects on the way readers with autism comprehended or memorised the meaning of a text; however, images did have an affect on the subjective perception of the ASD participants: they saw images as helpful cues to comprehension and memorisation. This result was unique to the ASD group (the control participants were mostly indifferent to the inclusion of images in text) and was consistent when measured through four differently-worded questions featuring in two different studies.

With regards to image type, visuals with strong and weak resemblance to their referents in reality (i.e., photographs and symbols) were fixated equally throughout the reading process, proving that both types of images are equally suitable to be used in texts for adults with autism. This may not be the case for children with autism though, as some evidence suggests that they develop symbolic understanding more slowly than typically developing peers (Allen 2009).

Another finding from the studies presented in Chapter 6 was that easy-to-read texts were well understood by adults with high-functioning autism but were not under-stimulating to them, in contrast to the perception of the control-group participants. The results suggest that when it comes to important information, easy-to-read texts are both preferred by and more

suitable for this group and are not perceived as under-stimulating.

Finally, in addition to the between-group differences in successfully answering the comprehension and memorisation questions after each text, there was a dramatic between-group difference in the answers to the memorisation questions. This finding suggests that, in comparison to the control group, the readers with autism struggled much more to memorise the information than to comprehend it. As a result, strategies aiming to aid comprehension in people with autism should also place a special emphasis on aiding memorisation, as memorisation may play a crucial role in the overall integration of information obtained from the text.

All the results from the studies above are integrated into a set of guidelines for accessible writing for adults with autism, presented in Section 6.6.3.

8.2.1 Impact

The findings from the studies of images and the guidelines for accessible writing that they were integrated into (Original Contribution 4) have the potential to change the way easy-to-read documents are developed. These guidelines are the first to be based on empirical evidence that describes what accessible documents for people with autism should look like in order to be optimally tailored to the needs of this population. The guidelines include sections on the insertion of images, on the types of images suitable for use, on the positioning of images, on different ways of supporting comprehension and memorisation, and on taking into account reading speed. Unlike previous

guidelines, we pay special emphasis on supporting not only comprehension but also memorisation. The reason for this is that for certain types of information (e.g. health and safety information, instructions for how to use a device, cooking recipes, etc.), memorisation (especially long-term retention) will be a crucial matter.

Many of the findings on the effects of images on attention and comprehension support previous research from the field of psychology by providing new evidence based on gaze data. As discussed in Chapter 2, attention differences between autistic and neurotypical individuals were described as early as the first mention of autism by Leo Kanner in 1943. This phenomenon had so far not been investigated in the context of reading and had not been supported by evidence based on gaze data. Furthermore, the attention differences in reading between the two groups of participants described in this study could influence the investigation of reading-comprehension deficits in people with autism from the perspective of attention differences (e.g., they focus on different aspects of the text and thus have reduced understanding of its meaning). The between-group differences in the memorisation of the information were even greater than the between-group differences in comprehension, which also suggests that at least some of the deficits in processing written information could be due to memory-related issues and thus it is not solely text simplification we should focus on but also reinforcement of the main information.

8.3 Processing of Web Pages

Research on the web-searching behaviour of people with autism is virtually non-existent, with the exception of a pilot study by Deering (2013) involving four participants with autism. As an initial step in this direction, in Chapter 7 we investigated the way adults with autism search for information within web pages, addressing research question five:

RQ5: Do web users with autism encounter barriers to finding information on web pages?

Our findings showed that web users with high-functioning autism do experience barriers when searching for information within web pages, as measured by their success in locating the required information on the pages. This finding suggests that, since differences in performance were registered for participants with only the mildest form of autism, web users with more severe forms of autism would be even more challenged when processing web pages. Furthermore, the web pages and tasks used in this study did not require reading of large chunks of text, which is why it can also be hypothesised that the processing of web pages with more text content (e.g. Wikipedia pages) could raise even greater barriers to web users with autism.

In terms of measuring the cognitive load associated with web-page processing, there were no statistically significant differences between the numbers of gaze fixations, or revisits or the average dwell times among the two

groups. Moreover, the results were inconclusive regarding the effects different levels of visual complexity may have on the success of finding information. Different levels of visual complexity of the web pages did not affect the performance of the two groups with the exception of the web pages with medium complexity. The reason for this result could be the fact that visual complexity was pre-defined using the ViCRAM algorithm (Michailidou 2010) rather than through human ratings, which is why it is uncertain whether the levels of complexity assigned to the web pages were inaccurate or whether visual complexity actually does not have an effect on finding information within web pages. A replication of the study with human pre-assessment of the complexity of the pages is thus highly recommended.

The survey questions used in this study pave the way for future exploration of web-searching behaviour in people with autism by exploring their subjective perceptions of how easy it is for them (in general) to find information on the web and to decide which of the returned links to open. The results revealed that even though the majority of both autistic and non-autistic participants felt that it was relatively easy for them to find information on the web, 16.66% of the participants with autism reported scores of "medium" or "very difficult".

8.3.1 Impact

The evidence from the study on web-page processing has the potential to effect, first and foremost, a very much needed shift in the web-accessibility

community towards the needs of people with cognitive disabilities: if even people with the mildest form of autism have difficulties processing web pages (Original Contribution 5), how accessible is the web for people at the lower ends of the spectrum? This result could be attributed to a multitude of reasons but when linked to results from the previous study on the effects of images on attention in people with autism, it could be discussed in the light of attention differences between people with autism and other groups of web users. This in turn challenges the perception that people with different types of cognitive disabilities have similar requirements when it comes to web-page accessibility, as is currently assumed in the WCAG 2.0 guidelines (Caldwell et al. 2008). One possible solution to this problem lies in the future personalisation of the web and, more specifically, in the area of adaptive user interfaces, which change according the needs of particular users based on log data. However, until that level of personalisation has been reached in mainstream use of the web, more studies into the user requirements of particular and well-defined user groups are needed in order to make the web accessible to everyone.

The next section discusses future work towards text and web accessibility for people with autism.

8.4 Future Work

There are many aspects to text and web accessibility. These aspects are equally important in making accessibility truly accessible for everyone.

In terms of text accessibility, future work is needed in order to evaluate how reliably gaze data can predict comprehension. Exploration of this relationship has the potential not only to strengthen the findings of this thesis, but with the emergence of cheap and unobtrusive eye-tracking technologies (e.g. in smart phones or tablets) it may prove a useful way of data collection in a real-world scenario.

Exploration of other extra-textual strategies for supporting text comprehension and memorisation is also a subject which may have a great impact on text accessibility for people with autism. In this thesis, we focused on images because previous research has shown the inclination of people to process information visually. However, other strategies such as the inclusion of mind maps, the highlighting of key words or the presentation of text in bullet-point form may prove a better way of supporting comprehension and memorisation. Furthermore, in spite of the fact that readers with autism struggle with pragmatics rather than with word decoding, the legibility of the text may also have impact on comprehension and memorisation. Exploration of the effects of font sizes, font types, line spacing and background colours may also prove a valuable avenue for future research.

While reading in autism has received a significant amount of attention

in the literature, web-searching behaviour in people with autism is one of the least explored areas in both accessibility research and in psychology. It is crucial that this issue be properly addressed by future research because of the crucial role the Internet has in the everyday lives of all people. In this thesis we have only scratched the surface of this problem by investigating information searching within web pages; much bigger questions remain unanswered, such as whether people with autism have difficulties defining queries, choosing the relevant links to open, coping with distractions, etc. Future work is needed to investigate the web-searching strategies of people with more severe forms of autism, as well as how they cope with web pages that contain a lot of textual information.

Finally, the most important aspect of accessibility is personalisation. No two people with autism are the same. Finding solutions towards the personalisation of comprehension cues, text presentation or web searching would probably be the biggest step towards truly accessible texts and web content and navigation. Last but not least, research towards personalisation is motivated by the well-known principle in accessibility that improved access for one user group can carry across to improved accessibility for everyone.

The 21st century is often referred to as "the age of information", and, indeed, access to information can change our lives in so many different ways. The role of accessibility research is to re-think constantly what "access" means. Thus, true text and web accessibility in the 21st century mean not just having access to information but also having access to its meaning.

Bibliography

Akpınar, M. E. & Yeşilada, Y. (2013*a*), Heuristic role detection of visual elements of web pages, *in* 'Web Engineering', Springer, pp. 123–131.

Akpınar, M. E. & Yesilada, Y. (2013*b*), Vision based page segmentation algorithm: Extended and perceived success, *in* 'Current Trends in Web Engineering', Springer, pp. 238–252.

Allen, M. L. (2009), 'Brief report: decoding representations: how children with autism understand drawings.', *Journal of autism and developmental disorders* **39**(3), 539–43.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/18810626*

Aluisio, S., Specia, L., Gasperin, C. & Scarton, C. (2010), Readability assessment for text simplification, *in* 'Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications', Association for Computational Linguistics, pp. 1–9.

American Psychiatric Association (2013), 'Diagnostic and Statistical Manual of Mental Disorders (5th ed.)'.

Anderson, J. (1983), 'Lix and rix: Variations on a little-known readability index', *Journal of Reading* **26**(6), 490–496.

Barzilay, R. & Elhadad, N. (2003), Sentence alignment for monolingual comparable corpora, *in* 'Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing', EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 25–32.
URL: *http://dx.doi.org/10.3115/1119355.1119359*

Bejerot, S., E. J. M. & Mörtberg, E. (2014), 'Social anxiety in adult autism spectrum disorder', *Psychiatry research* **220**((1-2)), 705–7.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/25200187*

Benjamin, R. G. (2011), 'Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty', *Educational Psychology Review* **24**(1), 63–88.
URL: *http://link.springer.com/10.1007/s10648-011-9181-8*

Bennetto, L., Pennington, B. F. & Rogers, S. J. (1996), 'Intact and Impaired Memory Functions in Autism', *Child Development* **67**(4), 1816–1835.

Bennöhr, J. (2005), A web-based personalised textfinder for language learners, Masters thesis, School of Informatics, University of Edinburgh. The Conference of the North American Chapter of the Association for Computational Linguistics (HLTNAACL-07).

Bialystok, E. (2000), 'Symbolic representation across domains in preschool children', *Journal of experimental child psychology* **76**(3), 173–89.
URL: *http://www.sciencedirect.com/science/article/pii/S0022096599925481*

Bormuth, J. R. (1967), 'Comparable cloze and multiple-choice comprehension test scores', *Journal of Reading* **10**(5), 291–299.

Bormuth, J. R. (1971), *Development of standards of readability: Toward a rational criterion of passage performance.*

Bosseler, A. & Massaro, D. W. (2003), 'Development and evaluation of computer-animated tutor for vocabulary and language learning in children with autism', *Journal of Autism and Developmental Disorders* **33**(6), 553–567.

Brega, A. G., Freedman, M. A., LeBlanc, W. G., Barnard, J., Mabachi, N. M., Cifuentes, M., Albright, K., Weiss, B. D., Brach, C. & West, D. R. (2015), 'Using the health literacy universal precautions toolkit to improve the quality of patient materials', *Journal of health communication* **20**(sup2), 69–76.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Britto, T. C. P. & Pizzolato, E. B. (2016), 'Towards web accessibility guidelines of interaction and interface design for people with autism spectrum disorder', In proceedings of ACHI 2016 : The Ninth International Conference on Advances in Computer-Human Interactions.

Britton, B. K. & Gülgöz, S. (1991), 'Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures.', *Journal of Educational Psychology* **83**(3), 329.

Brock, J., Norbury, C., Einav, S. & Nation, K. (2008), 'Do individuals with autism process words in context? evidence from language-mediated eye-movements', *Cognition* **108**(3), 896–904.

Bruce, B., Rubin, A. & Starr, K. (1981), 'Why readability formulas fail', *IEEE Transactions on Professional Communication* (1), 50–52.

Brugha, T., Cooper, S. A. & McManus, S. (2012), Estimating the prevalence of autism spectrum conditions in adults: Extending the 2007 adult psychiatric morbidity survey, Technical report, NHS, The Health and Social Care Information Centre, London.

Caldwell, B., Cooper, M., Reid, L. & Vanderheiden, G. (2008), 'Web Content Accessibility Guidelines 2.0 (WCAG 2.0)', W3C. `http://www.w3.org/TR/WCAG20/`.

Callan, J. & Eskenazi, M. (2007), Combining lexical and grammatical features to improve readability measures for first and second language texts, *in* 'Proceedings of NAACL HLT', pp. 460–467.

Caylor, J. S. et al. (1973), 'Methodologies for determining reading requirements of military occupational specialties.'.

Chall, J. S. & Dale, E. (1995), *Readability Revisited: the new Dale-Chall readability formula*, Brookline Books, Cambridge, Massachusetts.

Cohen, J. (1988), 'Statistical power analysis for the behavioral sciences. 2nd edn. hillsdale, new jersey: L'.

Coleman, E. B. (1971), *Developing a technology of written instruction: some determiners of the complexity of prose*, Teachers College Press, Columbia University, New York.

Collins-Thompson, K. & Callan, J. (2005), 'Predicting reading difficulty with statistical language models', *Journal of the American Society for Information Science and Technology* **56**(13), 1448–1462.

Collins-Thompson, K. & Callan, J. P. (2004), A language modeling approach to predicting reading difficulty., *in* 'HLT-NAACL', pp. 193–200.

Coltheart, M. (1981), 'The mrc psycholinguistic database', *The Quarterly Journal of Experimental Psychology Section A* **33**(4), 497–505.
URL: *http://dx.doi.org/10.1080/14640748108400805*

Dale, E. & Chall, J. S. (1948), 'A formula for predicting readability: Instructions', *Educational Research Bulletin* **27**(2), 37–54.

Dale, E. & Tyler, R. W. (1934), 'A study of the factors influencing the difficulty of reading materials for adults of limited reading ability', *The Library Quarterly: Information, Community, Policy* **4**(3), 384–412.

Dave, D. M. & Fernandez, J. M. (2015), 'Rising autism prevalence: Real or displacing other mental disorders? evidence from demand for auxiliary healthcare workers in california', *Economic Inquiry* **53**(1), 448–468.

Davis, F. B. (1946), 'Item-analysis data; their computation, interpretation, and use in test construction.', *Harvard Education Papers* .

Day, R. R. & Park, J.-S. (2005), 'Developing Reading Comprehension Questions', *Reading in a Foreign Language* **17**(1).

Deering, H. J. (2013), Opportunity for success: Website evaluation and scanning by students with autism spectrum disorders, Master's thesis, Iowa State University.

Dell'Orletta, F., Montemagni, S. & Venturi, G. (2011), Read-it: Assessing readability of italian texts with a view to text simplification, *in* 'Proceedings of the second workshop on speech and language processing for assistive technologies', Association for Computational Linguistics, pp. 73–83.

DeLoache, J. S. (2008), 'Symbolic Functioning in Very Young Children: Understanding of Pictures and Models', *Child Development* **62**(4), 736–752.

Dennis, M., Lazenby, A. L. & Lockyer, L. (2001), 'Inferential language in high-function children with autism', *Journal of autism and developmental disorders* **31**(1), 47–54.

Dettmer, S., Simpson, R. L., Myles, B. S. & Ganz, J. B. (2000), 'The use of visual supports to facilitate transitions of students with autism', *Focus on Autism and Other Developmental Disabilities* **15**(3), 163–169.

Dolch, E. W. (1948), *Problems in Reading*, The Garrard Press, Champain, IL.

Dornescu, I., Evans, R. & Orasan, C. (2013), A Tagging Approach to Identify Complex Constituents for Text Simplification, *in* 'Proceedings of Recent

Advances in Natural Language Processing', Hissar, Bulgaria, pp. 221 – 229.

DuBay, W. H. (2008), 'The principles of readability. 2004', *Costa Mesa: Impact Information* **76**.

Duchowski, A. (2009), *Eye Tracking Methodology: Theory and Practice*, second edn, Springer.

Dumais, S. T. (2004), 'Latent semantic analysis', *Annual review of information science and technology* **38**(1), 188–230.

Eden, G., Stein, J., Wood, H. & Wood, F. (1994), 'Differences in eye movements and reading problems in dyslexic and normal children', *Vision research* **34**(10), 1345–1358.

Ehrlich, S. F. & Rayner, K. (1981), 'Contextual effects on word perception and eye movements during reading', *Journal of verbal learning and verbal behavior* **20**(6), 641–655.

Eraslan, S., Yaneva, V., Yesilada, Y., Harper, S. & Mitkov, R. (n.d.), 'Web users with autism: Eye-tracking evidence of barriers and distractors', (Under Review).

Eraslan, S. & Yesilada, Y. (2015), 'Patterns in eyetracking scanpaths and the affecting factors', *Journal of Web Engineering* **14**(5-6), 363–385.

Evans, R., Orasan, C. & Dornescu, I. (2014), An evaluation of syntactic simplification rules for people with autism, *in* 'Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)', pp. 131–140.

Fei-Fei, L. & Russakovsky, O. (2013), 'Analysis of large-scale visual recognition', Bay Area Vision Meeting.

Feng, L. (2009), 'Automatic readability assessment for people with intellectual disabilities', *SIGACCESS Access. Comput.* (93), 84–91.
URL: *http://doi.acm.org/10.1145/1531930.1531940*

Feng, L., Elhadad, N. & Huenerfauth, M. (2009), Cognitively motivated features for readability assessment, *in* 'EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009', pp. 229–237.
URL: *http://www.aclweb.org/anthology/E09-1027*

Feng, L., Jansche, M., Huenerfauth, M. & Elhadad, N. (2010), A comparison of features for automatic readability assessment, *in* 'Proceedings of the 23rd International Conference on Computational Linguistics: Posters', COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 276–284.
URL: *http://dl.acm.org/citation.cfm?id=1944566.1944598*

Flesch, R. (1948), 'A new readability yardstick.', *Journal of applied psychology* **32**(3), 221.

Flesch, R. (1949), *The art of readable writing*, Harper, New York.

François, T. & Fairon, C. (2012), An ai readability formula for french as a foreign language, *in* 'Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning', Association for Computational Linguistics, pp. 466–477.

Frank, E. & Witten, I. H. (1998), Generating accurate rule sets without global optimization, *in* J. Shavlik, ed., 'Fifteenth International Conference on Machine Learning', Morgan Kaufmann, pp. 144–151.

Freyhoff, G., Hess, G., Kerr, L., Tronbacke, B. & Veken, K. V. D. (1998), Make it simple. european guidelines for the production of easy-to-read information for people with learning disability, Technical report, ILSMH European Association.

Friedman, M. G. & Bryen, D. N. (2007), 'Web accessibility design recommendations for people with cognitive disabilities', *Technology and Disability* **19**(4), 205–212.

Frith, U. (2003), *Autism. Explaining the enigma*, second edn, Blackwell Publishing, Oxford, UK.

Frith, U. & Snowling, M. (1983), 'Reading for meaning and reading for sound in autistic and dyslexic children', *Journal of Developmental Psychology* **1**, 329–342.

Fritz, C. O., Morris, P. E. & Richler, J. J. (2012), 'Effect size estimates: current use, calculations, and interpretation.', *Journal of Experimental Psychology: General* **141**(1), 2.

Fry, E. (1968), 'A readability formula that saves time', *Journal of reading* pp. 513–578.

Fry, E. (2004), *1000 Instant Words: The Most Common Words for Teaching Reading, Writing and Spelling*, Teacher Created Resources.

Gazepoint (2015), 'GP3 Gazepoint Eye tracker'.
URL: *http://www.eyegaze.com/instrument-specifications/*

George-Nektarios, T. (2013), 'Weka classifiers summary', *Athens University of Economics and Bussiness Intracom-Telecom, Athens* .

Gibson, E. (1998), 'Linguistic complexity: Locality of syntactic dependencies', *Cognition* **68**(1), 1–76.

Gilliland, J. (1972), 'Readability', *London: University of London Press Ltd.* .

Gottron, T. & Martin, L. (2009), Estimating web site readability using con-

tent extraction, *in* 'Proceedings of the 18th international conference on World wide web', ACM, pp. 1169–1170.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004), 'Coh-metrix: Analysis of text on cohesion and language.', *Behavioral Research Methods, Instruments, and Computers* **36**, 193–202.

Grandin, T. (2009), 'How does visual thinking work in the mind of a person with autism? a personal account', *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**(1522), 1437–1442.

Gronlund, N. E. (1982), *Constructing achievement tests*, Prentice Hall.

Gunning, R. (1952), *The technique of clear writing*, McGraw-Hill, New York.

Hadwin, J., Baron-Cohen, S., Howlin, P. & Hill, K. (1997), 'Does teaching theory of mind have an effect on the ability to develop conversation in children with autism?', *Journal of autism and developmental disorders* **27**(5), 519–537.

Halliday, M. A. & Hasan, R. (1976), 'Cohesion in', *English. Longman, London* .

Happe, F. (1997), 'Central coherence and theory of mind in autism: Reading homographs in context', *British Journal of Developmental Psychology* **15**, 1–12.

Happé, F. & Frith, U. (2006), 'The weak coherence account: Detail focused cognitive style in autism spectrum disorder', *Journal of Autism and Developmental Disorders* **36**, 5–25.

Happé, F. G. (1995), 'The role of age and verbal ability in the theory of mind task performance of subjects with autism', *Child development* **66**(3), 843–855.

Harper, S. & Yesilada, Y. (2008), Web accessibility and guidelines, *in* S. Harper & Y. Yesilada, eds, 'Web Accessibility: A Foundation for Research', 1st edn, Human-Computer Interaction Series, Springer, London, chapter 6, pp. 61–78.

Harris, T. L. & Hodges, R. E. (1995), *The Literacy Dictionary: The Vocabulary of Reading and Writing*, International Reading Association.

Hartley, C. & Allen, M. L. (2014), 'Intentions vs. resemblance: understanding pictures in typical development and autism.', *Cognition* **131**(1), 44–59.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/24440433*

Hastie, T., Tibshirani, R. & Friedman, J. (2001), 'The elements of statistical learning. 2001', *NY Springer* .

Heilman, M., Collins-Thompson, K. & Eskenazi, M. (2008), An analysis of statistical models and features for reading difficulty prediction, *in* 'Proceedings of the Third Workshop on Innovative Use of NLP for Build-

ing Educational Applications', Association for Computational Linguistics, pp. 71–79.

Hobson, R. P. (2012), 'Autism, literal language and concrete thinking: Some developmental considerations', *Metaphor and Symbol* **27**(1), 4–21.

Hornof, A. J. & Halverson, T. (2002), 'Cleaning up systematic error in eye-tracking data by using required fixation locations', *Behavior Research Methods, Instruments, & Computers* **34**(4), 592–604.

Hovy, D., Srivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Sanders, W. & Hovy, E. (2013), Identifying metaphorical word use with tree kernels, *in* 'Proceedings of the First Workshop on Metaphor in NLP', Citeseer, pp. 52–57.

Hull, L. (1979), 'Measuring the readability of technical writing', Proceedings of the 26th International Technical Communications Conference, Los Angeles.

IBM Corp. (2011), 'IBM SPSS Statistics for Windows, Version 20.0'.

Inui, K., Yamamoto, S. & Inui, H. (2001), Corpus-based acquisition of sentence readability ranking models for deaf people., *in* 'NLPRS', pp. 159–166.

Ivory, M. Y., Sinha, R. R. & Hearst, M. A. (2001), Empirically validated web page design metrics, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, pp. 53–60.

Jarrold, C. & Brock, J. (2004), 'To match or not to match ? methodological issues in autism related research', *Journal of autism and developmental disorders* **34**(1), 81–86.

Joachims, T. (1998), Text categorization with support vector machines: Learning with many relevant features, *in* 'European conference on machine learning', Springer, pp. 137–142.

Jordanova, V., Evans, R. & Cerga-Pashoja, A. (2013), FIRST Deliverable - Benchmark report (result of piloting task), Technical Report D7.2, Central and Northwest London NHS Foundation Trust, London, UK.
URL: *http://first-asd.eu/?q=system/files/FIRST$_D$7.2$_2$0140414.pdf*

Just, M. A. & Carpenter, P. A. (1980), 'A theory of reading: from eye fixations to comprehension.', *Psychological review* **87**(4), 329.

Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J. & Just, M. A. (2006), 'Sentence comprehension in autism: thinking in pictures with decreased functional connectivity', *Brain* **129**(9), 2484–2493.

Kane, L., Carthy, J. & Dunnion, J. (2006), Readability applied to information retrieval, *in* 'European Conference on Information Retrieval', Springer, pp. 523–526.

Kanner, L. (1943), 'Autistic Disturbances of Affective Contact', *Nervous Child* **2**, 217–250.

Kanungo, T. & Orr, D. (2009), Predicting the readability of short web summaries, *in* 'Proceedings of the Second ACM International Conference on Web Search and Data Mining', ACM, pp. 202–211.

Kennedy, A., Hill, R. & Pynte, J. (2003), 'The dundee corpus', *Proceedings of the 12th European conference on eye movement.* .

Kennedy, A., Pynte, J., Murray, W. S. & Paul, S.-A. (2013), 'Frequency and predictability effects in the dundee corpus: An eye movement analysis', *The Quarterly Journal of Experimental Psychology* **66**(3), 601–618.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. (1975), Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Technical report, CNTECHTRA Research Branch Report.

Kintsch, W. & van Dijk, T. (1978), 'Toward a model of text comprehension and production', *Psychological Review* **85**(5), 363–394.

Kintsch, W. & Vipond, D. (1977), 'Reading comprehension and readability in educational practice and psychological theory', *In Lars-Goran Nilsson (Ed.),Proceedings of the Conference on Memory.* .

Kispal, A. (2008), 'Effective teaching of inference skills for reading: Literature review'.

Klein, D. & Manning, C. D. (2003), Natural language parsing, *in* 'Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference', Vol. 15, MIT Press, p. 3.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004), 'Length, frequency,

and predictability effects of words on eye movements in reading', *European Journal of Cognitive Psychology* **16**(1-2), 262–284.

Kliegl, R., Nuthmann, A. & Engbert, R. (2006), 'Tracking the mind during reading: the influence of past, present, and future words on fixation durations.', *Journal of experimental psychology: General* **135**(1), 12.

Lakoff, G. & Johnson, M. (2003), 'Metaphors we live by. 1980', *Chicago: U of Chicago P* .

Landauer, T. K., Foltz, P. W. & Laham, D. (1998), 'An introduction to latent semantic analysis', *Discourse processes* **25**(2-3), 259–284.

Laufer, B. & Nation, P. (1999), 'A vocabulary-size test of controlled productive ability', *Language testing* **16**(1), 33–51.

Lively, B. A. & Pressey, S. L. (1923), *A method for measuring the" vocabulary burden" of textbooks.*

Lorge, I. (1944), 'Predicting readability.', *Teachers College Record* .

Lovaas, O. I. & Schreibman, L. (1971), 'Stimulus Overselectivity of Autistic Children in a Two Stimulus Situation', *Behavior Research and Therapy* **9**, 305–310.

MacKay, G. & Shaw, A. (2004), 'A comparative study of figurative language in children with autistic spectrum disorders.', *Child Language Teaching and Therapy* **20**(13).

Martos, J., Freire, S., González, A., Gil, D., Evans, R., Jordanova, V., Cerga, A., Shishkova, A. & Orasan, C. (2013), FIRST Deliverable - User preferences: Updated, Technical Report D2.2, Deletrea, Madrid, Spain.
URL: *http://first-asd.eu/?q=system/files/FIRST$_D$2.2$_2$0130531.pdf*

Max, A. (2000), Syntactic simplification - an application to text for aphasic readers, Mphil in computer speech and language processing, University of Cambridge, Wolfson College.

McLaughlin, H. G. (1969), 'SMOG grading - a new readability formula', *Journal of Reading* pp. 639–646.

McNamara, D. S., Graesser, A. C., McCarthy, P. M. & Cai, Z. (2014), *Automated evaluation of text and discourse with Coh-Metrix*, Cambridge University Press.

McNamara, D. S. & Kintsch, W. (1996), 'Learning from texts: Effects of prior knowledge and text coherence', *Discourse processes* **22**(3), 247–288.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M. & Graesser, A. C. (2010), 'Coh-Metrix: Capturing Linguistic Features of Cohesion'.
URL: *http://www.tandfonline.com/doi/abs/10.1080/0163853090295994 3*

Michailidou, E. (2006), 'Vicram: visual complexity rankings and accessibility metrics', *ACM SIGACCESS Accessibility and Computing* (86), 24–27.

Michailidou, E. (2010), Visual complexity rankings and accessibility metrics, PhD thesis, The University of Manchester.

Michailidou, E., Harper, S. & Bechhofer, S. (2008), Visual complexity and aesthetic perception of web pages, *in* 'Proceedings of the 26th annual ACM international conference on Design of communication', ACM, pp. 215–224.

Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Miller, G. A., Beckwith, R., Fellbaum, C. & Gross, D. (1990), 'Wordnet: An online lexical database', *International Journal of Lexicography* **3**(4), 235244.

Mitkov, R. (2002), *Anaphora Resolution*, Longman, Harlow, Essex.

Mohler, M., Rink, B., Bracewell, D. B. & Tomlinson, M. T. (2014), A novel distributional approach to multilingual conceptual metaphor recognition., *in* 'COLING', pp. 1752–1763.

Nation, K., Adams, J. W., Bowyer-Crane, C. A. & Snowling, M. J. (1999), 'Working memory deficits in poor comprehenders reflect underlying language impairments', *Journal of experimental child psychology* **73**(2), 139–158.

Nation, K., Clarke, P., Wright, B. & Williams, C. (2006), 'Patterns of reading ability in children with autism-spectrum disorder', *Journal of Autism & Developmental Disorders* **36**, 911–919.

Newbold, N. & Gillam, L. (2010), The linguistics of readability: The next

step for word processing, *in* 'Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids', Association for Computational Linguistics, pp. 65–72.

Nickerson, C. A. & Cartwright, D. S. (1984), 'The university of colorado meaning norms', *Behavior Research Methods, Instruments, & Computers* **16**(4), 355–382.

Niculae, V. & Yaneva, V. (2013), Computational considerations of comparisons and similes., *in* 'ACL (Student Research Workshop)', pp. 89–95.

Nomura, M., Nielsen, G. S. & Tronbacke, B. (2010), Guidelines for easy-to-read materials/rev, Technical report, International Federation of Library Associations and Institutions, IFLA Headquarters, The Hague.

Norbury, C. F. (2014), 'Atypical pragmatic development', *Pragmatic Development in First Language Acquisition* **10**, 343.

Nuttall, C. (1996), *Teaching reading skills in a foreign language*, ERIC.

O'Connor, I. M. & Klein, P. D. (2004), 'Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders', *Journal of autism and developmental disorders* **34**(2), 115–127.

Oliver, S. (1998), *Understanding Autism*, Oxford Brookes University, UK.

Orasan, C., Evans, R. & Dornescu, I. (2013), 'Towards multilingual europe 2020: A romanian perspective, chapter text simplification for people with autistic spectrum disorders'.

Ozasa, T., Weir, G. & Fukui, M. (2007), Measuring readability for japanese learners of english, *in* 'Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics', pp. 122–125.

Ozasa, T., Weir, G. & Fukui, M. (2008), Toward a readability index for japanese learners of efl, *in* 'Proceedings of the 13th Conference of Pan-Pacific Association of Applied Linguistics (PAAL08). University of Hawaii, Manoa: Pan-Pacific Association of Applied Linguistics. Available from http://www. cis. strath. ac. uk/cis/research/publications/papers/strath cis publication', Vol. 2263.

Paivio, A., Smythe, P. C. & Yuille, J. C. (1968), 'Imagery versus meaningfulness of nouns in paired-associate learning.', *Canadian Journal of Psychology/Revue canadienne de psychologie* **22**(6), 427.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K. & Newman, J. K. (2004), The determinants of web page viewing behavior: an eye-tracking study, *in* 'Proceedings of the 2004 symposium on Eye tracking research & applications', ACM, pp. 147–154.

Pastor, G. C., Mitkov, R., Afzal, N. & Pekar, V. (2008), Translation univer-

sals: do they exist? a corpus-based nlp study of convergence and simplification, *in* '8th AMTA conference', pp. 75–81.

Pavlov, N. (2014), 'User Interface for People with Autism Spectrum Disorders', **2014**(February), 128–134.

Pearson, P. D. & Johnson, D. D. (1978), *Teaching reading comprehension*, Harcourt School.

Perfetti, C. A., Landi, N. & Oakhill, J. (2005), 'The acquisition of reading comprehension skill.'.

Petersen, S. E. & Ostendorf, M. (2007), Text simplification for language learners: a corpus analysis., *in* 'SLaTE', Citeseer, pp. 69–72.

Pikulski, J. J. (2002), 'Readability'.

Pilán, I., Volodina, E. & Johansson, R. (2014), Rule-based and machine learning approaches for second language sentence-level readability, *in* 'Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications', pp. 174–184.

Plimpton, S. & Root, J. (1994), 'Materials and strategies that work in low literacy health communication.', *Public health reports* **109**(1), 86.

Ploog, B. O. (2010), 'Stimulus overselectivity four decades later: A review of the literature and its implications for current research in autism spectrum

disorder', *Journal of autism and developmental disorders* **40**(11), 1332–1349.

Putnam, C. & Chong, L. (2008), Software and technologies designed for people with autism: What do users want?, *in* 'Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility', Assets '08, ACM, New York, NY, USA, pp. 3–10.
URL: *http://doi.acm.org/10.1145/1414471.1414475*

Quill, K. A. (1997), 'Instructional Considerations for Young Children with Autism: The Rationale for Visually Cued Instruction', *Journal of Autism and Developmental Disorders* **27**(6).
URL: *http://link.springer.com/article/10.1023/A:1025806900162#page-2*

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine learning* **1**(1), 81–106.

Quinlan, J. R. (1987), 'Simplifying decision trees', *International journal of man-machine studies* **27**(3), 221–234.

Rayner, K. (1975), 'The perceptual span and peripheral cues in reading', *Cognitive Psychology* **7**(1), 65–81.

Rayner, K. (1998), 'Eye movements in reading and information processing: 20 years of research.', *Psychological bulletin* **124**(3), 372.

Rayner, K. (2009), 'Eye movements and attention in reading, scene perception, and visual search', *The quarterly journal of experimental psychology* **62**(8), 1457–1506.

Rayner, K. & Duffy, S. A. (1986), 'Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity', *Memory & Cognition* **14**(3), 191–201.

Rayner, K., Pollatsek, A., Ashby, J. & Clifton Jr, C. (2012), *Psychology of reading*, Psychology Press.

Reichle, E. D., Rayner, K. & Pollatsek, A. (1999), 'Eye movement control in reading: Accounting for initial fixation locations and refixations within the ez reader model', *Vision research* **39**(26), 4403–4411.

Reichle, E. D., Rayner, K. & Pollatsek, A. (2003), 'The ez reader model of eye-movement control in reading: Comparisons to other models', *Behavioral and brain sciences* **26**(04), 445–476.

Rello, L., Baeza-Yates, R., Bott, S. & Saggion, H. (2013), Simplify or help?: text simplification strategies for people with dyslexia, *in* 'Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility', ACM, p. 15.

Rello, L., Baeza-Yates, R., Dempere-Marco, L. & Saggion, H. (2013), Frequent words improve readability and short words improve understandabil-

ity for people with dyslexia, *in* 'IFIP Conference on Human-Computer Interaction', Springer, pp. 203–219.

Rello, L. & Ballesteros, M. (2015), Detecting readers with dyslexia using machine learning with eye tracking measures, *in* 'Proceedings of the 12th Web for All Conference', W4A '15, ACM, New York, NY, USA, pp. 16:1–16:8.

URL: *http://doi.acm.org/10.1145/2745555.2746644*

Rello, L., Saggion, H., Baeza-Yates, R. & Graells, E. (2012), Graphical schemes may improve readability but not understandability for people with dyslexia, *in* 'Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations', Association for Computational Linguistics, pp. 25–32.

Root, J. & Stableford, S. (1999), 'Easy-to-read consumer communications: a missing link in medicaid managed care', *Journal of Health Politics, Policy and Law* **24**(1), 1–26.

Rosenfeld, R. (2000), 'Two decades of statistical language modeling: Where do we go from here?'.

Rundblad, G. & Annaz, D. (2010), 'The atypical development of metaphor and metonymy comprehension in children with autism', *Autism* **14**(1), 29–46.

Russell, J., Jarrold, C. & Henry, L. (1996), 'Working memory in children with

autism and with moderate learning difficulties', *Journal of Child Psychology and Psychiatry* **37**(6), 673–686.

Saldaña, D. & Frith, U. (2007), 'Do readers with autism make bridging inferences from world knowledge?', *Journal of Experimental Child Psychology* **96**(4), 310–319.

Sampath, H., Sivaswamy, J. & Indurkhya, B. (2010), Assistive systems for children with dyslexia and autism, *in* 'SIGACCESS Newsletter'.

Sampath, Harini, S. J. I. B. (2010), Assistive Systems for Children with Dyslexia and Autism, *in* 'SIGACCESS Newsletter', number 96, pp. 32–36.

Sansosti, F. J., Was, C., Rawson, K. A. & Remaklus, B. L. (2013), 'Eye movements during processing of text requiring bridging inferences in adolescents with higher functioning autism spectrum disorders: A preliminary investigation', *Research in Autism Spectrum Disorders* **7**(12), 1535–1542.

Sasson, N. J. & Elison, J. T. (2012), 'Eye tracking young children with autism', *JoVE (Journal of Visualized Experiments)* (61), e3675–e3675.

Schwarm, S. E. & Ostendorf, M. (2005), Reading level assessment using support vector machines and statistical language models, *in* 'Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, pp. 523–530.

Senter, R. J. & Smith, E. A. (1967), Automated Readability Index, Technical

Report AMRL-TR-6620, Wright-Patterson Air Force Base.

URL: *http://www.dtic.mil/cgi-bin/GetTRDoc?AD=AD0667273*

Shalev-Shwartz, S., Singer, Y., Srebro, N. & Cotter, A. (2011), 'Pegasos: Primal estimated sub-gradient solver for svm', *Mathematical programming* **127**(1), 3–30.

Shutova, E. (2010), Models of metaphor in nlp, *in* 'Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics', ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 688–697.

URL: *http://dl.acm.org/citation.cfm?id=1858681.1858752*

Si, L. & Callan, J. (2001), A statistical model for scientific readability, *in* 'Proceedings of the Tenth International Conference on Information and Knowledge Management', CIKM '01, ACM, New York, NY, USA, pp. 574–576.

URL: *http://doi.acm.org/10.1145/502585.502695*

Siddharthan, A. (2004), Syntactic Simplification and Text Cohesion, PhD thesis, University of Cambridge.

Slatin, J. M. & Rush, S. (2003), *Maximum Accessibility: Making Your Website More Usable for Everyone*, Addison-Wesley, Boston.

Smith, D. R., Stenner, A. J., Horabin, I. & Malbert Smith, I. (1989), The lexile scale in theory and practice: Final report., Technical report, MetaMet-

rics (ERIC Document Reproduction Service No. ED307577)., Washington, DC:.

Speirs, S., Yelland, G., Rinehart, N. & Tonge, B. (2011), 'Lexical processing in individuals with high-functioning autism and asperger's disorder', *Autism* p. 1362361310386501.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A. & Krennmayr, T. (2010), 'Vu amsterdam metaphor corpus'.

Stymne, S., Tiedemann, J., Hardmeier, C. & Nivre, J. (2013), Statistical machine translation with readability constraints, *in* 'Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16', number 085, Linköping University Electronic Press, pp. 375–386.

Taylor, W. L. (1953), 'Cloze procedure: a new tool for measuring readability.', *Journalism and Mass Communication Quarterly* **30**(4), 415.

Tronbacke, B. (1997), Guidelines for Easy-to-Read Materials, Technical report, IFLA, The Hague.

Vajjala Balakrishna, S. (2015), Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications, PhD thesis, Universität Tübingen.

Vajjala, S. & Meurers, D. (2012), On improving the accuracy of readability

classification using insights from second language acquisition, *in* 'In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 163–173.

Vajjala, S. & Meurers, D. (2014), 'Readability assessment for text simplification: From analysing documents to identifying sentential simplifications', *ITL-International Journal of Applied Linguistics* **165**(2), 194–222.

Vogel, M. & Washburne, C. (1928), 'An objective method of determining grade placement of children's reading material', *The Elementary School Journal* **28**(5), 373–381.

ΩŠtajner et al.

Štajner, S., Mitkov, R. & Pastor, G. C. (2014), *Simple or not simple? A readability question*, Springer-Verlag, Berlin.

Wan, X., Li, H. & Xiao, J. (2010), Eusum: extracting easy-to-understand english summaries for non-native readers, *in* 'Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 491–498.

Whyte, E. M., Nelson, K. E. & Khan, K. S. (2013), 'Learning of idiomatic language expressions in a group intervention for children with autism', *Autism* **17**(4), 449–464.

Whyte, E. M., Nelson, K. E. & Scherf, K. S. (2014), 'Idiom, syntax, and

advanced theory of mind abilities in children with autism spectrum disorders', *Journal of Speech, Language, and Hearing Research* **57**(1), 120–130.

Williams, D. L., Goldstein, G., Carpenter, P. A. & Minshew, N. J. (2005), 'Verbal and spatial working memory in autism', *Journal of autism and developmental disorders* **35**(6), 747–756.

Witten, I. H. & Frank, E. (2005), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Yan, X., Song, D. & Li, X. (2006), Concept-based document readability in domain specific information retrieval, *in* 'Proceedings of the 15th ACM international conference on Information and knowledge management', ACM, pp. 540–549.

Yaneva, V. (2015), Easy-read documents as a gold standard for evaluation of text simplification output, *in* 'Proceedings of the Student Research Workshop', INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pp. 30–36.
URL: *http://www.aclweb.org/anthology/R15-2005*

Yaneva, V. & Evans, R. (2015), Six good predictors of autistic text comprehension, *in* 'Proceedings of the International Conference Recent Advances in Natural Language Processing', INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pp. 697–706.
URL: *http://www.aclweb.org/anthology/R15-1089*

Yaneva, V., Mitkov, R., Orasan, C. & Temnikova, I. (n.d.), 'Can image insertion aid reading comprehension and memorisation in adults with autism? user studies, a prototype and guidelines.', (Under Review).

Yaneva, V., Temnikova, I. & Mitkov, R. (2015), Accessible texts for autism: An eye-tracking study, *in* 'Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility', ASSETS '15, ACM, New York, NY, USA, pp. 49–57.
URL: *http://doi.acm.org/10.1145/2700648.2809852*

Yaneva, V., Temnikova, I. & Mitkov, R. (2016), 'A corpus of text data and gaze fixations from autistic and non-autistic adults.', Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia,.